TEXT COMPLEXITY AND TEXT SIMPLIFICATION IN THE CRISIS MANAGEMENT DOMAIN

Irina Temnikova B.A., M.A.

A thesis submitted in partial fulfilment of the
requirements of the University of Wolverhampton
for the degree of Doctor of Philosophy

April 2012

Signature.....

Date.....

Abstract

Due to the fact that emergency situations can lead to substantial losses, both financial and in terms of human lives, it is essential that texts used in a crisis situation be clearly understandable.

This thesis is concerned with the study of the complexity of the crisis management sub-language and with methods to produce new, clear texts and to rewrite pre-existing crisis management documents which are too complex to be understood. By doing this, this interdisciplinary study makes several contributions to the crisis management field. First, it contributes to the knowledge of the complexity of the texts used in the domain, by analysing the presence of a set of written language complexity issues derived from the psycholinguistic literature in a novel corpus of crisis management documents. Second, since the text complexity analysis shows that crisis management documents indeed exhibit high numbers of text complexity issues, the thesis adapts to the English language controlled language writing guidelines which, when applied to the crisis management language, reduce its complexity and ambiguity, leading to clear text documents. Third, since low quality of communication can have fatal consequences in emergency situations, the proposed controlled language guidelines and a set of texts which were re-written according to them are evaluated from multiple points of view. In order to achieve that, the thesis both applies existing evaluation approaches and develops new methods which are more appropriate for the task. These are used in two evaluation experiments – evaluation on extrinsic tasks and evaluation of users' acceptability.

The evaluations on extrinsic tasks (evaluating the impact of the controlled language on text complexity, reading comprehension under stress, manual translation, and machine translation tasks)

show a positive impact of the controlled language on simplified documents and thus ensure the quality of the resource. The evaluation of users' acceptability contributes additional findings about manual simplification and helps to determine directions for future implementation.

The thesis also gives insight into reading comprehension, machine translation, and cross-language adaptability, and provides original contributions to machine translation, controlled languages, and natural language generation evaluation techniques, which make it valuable for several scientific fields, including Linguistics, Psycholinguistics, and a number of different sub-fields of NLP.

Dedication and Acknowledgements

The greatest good you can do for another is not just to share your riches but to reveal to him his own. (Benjamin Disraeli)

This thesis is dedicated to my parents, the nuclear physicists Prof. Dr. Galina Maneva and Prof. Dr. Petar Temnikov, to whom I am deeply grateful for having inspired me to enter the scientific profession and for their extensive moral and scientific support in times that were difficult for me.

I gratefully acknowledge the support of many people who were important to me in the conduct of the research reported in this thesis. Their names and the reasons that I am grateful to them follow below.

First of all, I would like to express my gratitude to my Director of Studies, Prof. Dr. Ruslan Mitkov, for having allowed me to fulfil my dream of working in the computational linguistics field, for having exposed me to extensive computational linguistics knowledge thanks to my work for the Journal of Natural Language Engineering and the active scientific environment in his group, and for leading me through my thesis. Secondly, I would like to express my gratefulness to the other members of my supervisory team—Dr. Le An Ha and Mr. Richard Evans—for their constant availability and for giving me valuable lessons in structuring and clearly expressing the ideas in my thesis. Their contributions to my research were invaluable.

Next, a very big “thank you!” goes to all of the great, experienced researchers who helped me with lots of useful discussions and clever ideas during the course of my thesis, such as: Dr. Constantin Orasan, Dr. Anke Buttner, Prof. Dr. Albena Vassileva, Dr. Lucia Specia, and Dr. Sobha Lalitha Devi.

I would like also to express a big “thank you” to Prof. Dr. Galia Angelova, for having given me the precious opportunity to come into close touch with the field of computational linguistics by involving me in one of the most prestigious NLP conferences—RANLP—and for having shown me support during the whole of my Ph.D. studies.

Special thanks go to Dr. Ralf Steinberger for having given me a strong push in the beginning of my career as a computational linguistics trainee, and to Dr. Katerina Gachevska, who gave me insights regarding how to organize the effective writing of my thesis and many other useful things.

I would also like to express my enormous gratitude to a very special person, Dr. Kevin Bretonnel Cohen, for having had the patience and perseverance to read, proofread, and provide useful scientific comments about all of the chapters of my dissertation and for having urged me not to give up and to keep writing. I would also like to thank Alison Carminke and Erin Stokes for proofreading several of my papers that contributed to this thesis.

Additionally, I would like to thank very much the members of my research group for their precious ideas before my viva, and particularly specific colleagues and fellow Ph.D. students for their availability and their support during the course of my thesis, such as Iustina Ilisei, Iustin Dornescu, Andrea Varga, Dr. Raphael Rubino (in particular for his programming help with Python scripts), Ivelina Nikolova, Natalia Konstantinova (for her support in the Journal of Natural Language Engineering work), Dr. Georgiana Puscasu, and Dr. Laura Hasler. Special thanks also go to Ruslana

Margova, who always replied promptly and with enthusiasm to ideas for our shared research. I would also like to mention some important people from my past: my grandmother, Dr. Cvetana Maneva; my grandfather, Mihail Manev; and my elementary school teacher, Marina Genina, and my favorite high school teacher, Prof. Tonina Fancello, for having shown me the nice sides of knowledge and for inspiring me to follow that light. Last but not least, I would like to thank my ex-boyfriend Mihail Milushev, for having helped me fight my fear of programming by showing me how interesting Perl is and for having taught me to program in Perl.

And finally, I would like to say a big “thank you” to all of my friends and my brother for the patience, understanding, friendship, and support that they have given me. They urged me not to give up and to “keep pushing” for all these years. You know who you are. THANK YOU!

Table of Contents

List of Abbreviations

AECMA – Association Européenne des Constructeurs de Matériel Aérospatial

BNC – British National Corpus

CALP – Computer-Aided Language Processing

CAT – Computer-Aided Translation

CL – Controlled Language

CLCM – Controlled Language for Crisis Management

CM – Crisis Management

CMC – Crisis Management Corpus

CNA – Choice-Network Analysis

CNL – Controlled Natural Language

f-ATS – fully-Automatic Text Simplification systems

HTML – HyperText Markup Language

IGP – Instructions for the General Population

ManT – Manual Translation

MCQ – Multiple-Choice Questions

MT – Machine Translation

Mtranslatability – Machine Translatability

NLG – Natural Language Generation

NLP – Natural Language Processing

NLTK – Natural Language Toolkit

NP – Noun Phrase

PE – Post-Editing

PP – Prepositional Phrase

TC – Text Complexity

TM – Translation Memories

TS – Text Simplification

XML – eXtensible Markup Language

Chapter 1 – Introduction

1.1. Context and Motivations

Crisis Management (CM, also called *Disaster Management*) is an aspect of risk management (Regester and Larkin 2005). It can be defined as the management of a dangerous situation which has already happened, ranging from a man-made disaster (such as terrorism or an unintentional technological breakdown) to a natural disaster (such as an earthquake, a tornado, or a fire), which can occur to "individuals, companies and countries" and can lead to "a substantial loss of life, money, assets, and productivity" (Schneid and Collins 2001).

Due to the recently increasing number of emergency situations with severe consequences (such as the attacks to the World Trade Center[1] or the Haiti earthquake[2]), the attention to the Crisis Management field is currently increasing (Coppola 2007). It is known that during the initial steps of disaster preparedness and prevention (*disaster identification* and *disaster profiling,* Coppola 2007), disaster managers must identify every possible primary or secondary hazard (defined by Coppola 2007) as "a source of potential harm to a community") which could lead to a disaster. Communication is considered to be an important technological hazard, whose technologies, organisation, and resources must be kept under control, as it is known that under stress conditions, human comprehension is altered (Kiwan et al. 1999) due to the very short reaction time (Ogrizek and Guillery 1999; Winerman 2009). In fact, several deadly accidents which were due to flaws in communication management have already occurred:

1 http://www.nist.gov/el/disasterstudies/wtc/. Last accessed on March 27th, 2012.
2 http://www.washingtonpost.com/wp-dyn/content/article/2010/02/09/AR2010020904447.html. Last accessed on March 27th, 2012.

- The Tenerife air crash (Air Line Pilots Association, 1977) was the largest air crash in history. It happened in Tenerife in 1977 and took more than five hundred lives. The disaster was due to lack of a common language and misunderstanding between the Air Traffic Control Tower and the pilots of the two airplanes involved.

- The Pécrot rail crash[3] occurred in 2001 and is considered to be the worst Belgian train disaster in history. It was due to lack of a common language between the departing and arriving stations' signalmen, who belonged to the French and Flemish parts of Belgium, in which it is only required to speak one of the two languages.

- The Scandinavian Star ferry disaster (Solheim et al. 1992) occurred in 1990 and killed over 150 people. One of the reasons for the disaster was the fact that most of the cabin crew could not communicate with the passengers, due to not knowing any foreign languages.

- In addition, studies have shown that in 1998 incomprehensible instructions were among the causes of car accidents in the U.S.A., resulting in fatal injuries to children. In fact, it has been shown that over 80% of child seats were not used properly and that more than 90% of the child seats' instructions did not correspond to the reading level of average American citizens (DuBay 2004).

In order to avoid situations in which badly designed communication plays a negative role in crisis situations, it is of crucial importance that the messages expressed by CM documents be correctly comprehended and straightforward to understand (Seeger et al. 1998; Coppola 2007).

Although thousands of CM texts do already exist and more and more of them are currently being

3 http://news.bbc.co.uk/1/hi/world/europe/1251789.stm, last accessed on March 1st, 2011.

produced, the contribution of the Natural Language Processing (NLP) field and of Linguistics to the field is under-developed. The focus of most NLP approaches is on detecting crisis events in texts on the basis of linguistic cues and does not approach the matter of crisis communication efficiency. These approaches consist of applying information extraction to Twitter, open discussion forums, blogs, and online news articles (Corvey et al., 2010; Ireson, 2009; Mark, 2012; Steinberger et al., 2009). Applications in Biomedical NLP have instead focussed on detecting epidemics and thus conducting epidemic surveillance on the basis of automatically processed clinical notes (Chapman et al., 2005; Conway et al., 2009).

A few linguistic approaches, although still relevant to the NLP field, have addressed the complexity and ambiguity of the communication involved in crisis situations with the aim of ensuring high quality of the prevention, management, and response stages of CM. These approaches consist of a few controlled languages (described in more detail in Section 2.3.2) which created specifications for the documentation and protocols for communication in situations in which a crisis is imminent or already occurring.

The shortcoming of these communication-focussed approaches is that they focus either on specialised document types, such as aircraft documentation (AECMA, 1995), or on communication in a restricted set of crisis management situations, such as the cross-border communication of the Channel Tunnel security officials (Johnson, 1993), or on specific European languages, such as French, Spanish, and Polish (Renahy et al., 2010; Blanco, 2009; Rudas, 2009). No scientific study of the communication efficiency and the comprehensibility of documents used in crisis management situations for the English language has ever been conducted. This is additionally of concern as English is a globally used language across the world and many CM documents are created in it. Due to the fact that incorrectly transmitted communication in the crisis management domain can lead to

fatal consequences and to the lack of scientific studies investigating the comprehensibility and methods of simplification of crisis management documents in English, performing such an investigation is essential.

1.2. Aims, Hypotheses, and Contributions

This thesis provides original contributions to the study of the language complexity and comprehensibility of Crisis Management written documents, as well as to methods of rewriting them in simple and straightforward language. This section presents the aims and the original contributions of this thesis. Section 1.2.1 presents the aims of and the fundamental assumptions behind the research in this thesis, Section 1.2.2. specifies the goals to be achieved, and Section 1.2.3 presents the contributions of this thesis to the knowledge of the subject under focus.

1.2.1. Thesis aims and research hypotheses

This thesis has three main aims, as follows:

The first aim is to investigate the comprehensibility and the text complexity of crisis management documents written in English. This aim is motivated by the crucial importance of clearly and correctly transmitted communication in crisis situations and the lack of any such studies for English.

The second aim of this thesis is to propose a method of rewriting existing crisis management documents into clear and straightforward ones and of creating clear and straightforward CM

documents from scratch. This aim is motivated by the large number of existing documents and the exponential production of new crisis management documents world-wide, due to the fast development of the CM field.

Finally, **the third aim** of this thesis is to evaluate the impact of the proposed approaches on primary tasks (text comprehension) and secondary tasks (such as translation to other languages) important for the domain.

In order for these aims to be achieved, the investigation is based on the following research hypotheses, which will be tested further in the thesis:

- Documents written in the crisis management domain are formulated in a specific sub-language. This hypothesis will be tested through a linguistic analysis of crisis management documents.

- Text complexity phenomena which may decrease the comprehensibility of written documents can be automatically evaluated quantitatively on the basis of linguistic cues. This hypothesis will be tested by studying the existing work in measuring text complexity.

- The approach to measurement of the text complexity of documents in the CM domain, the appropriate method for rewriting emergency documents into clear forms or creating new documents from scratch, as well as the evaluation approaches employed to measure the impact of the proposed methods must be tailored to the domain, due to the specificity of the sub-language and the circumstances in which these documents are used (emergency situations). This hypothesis will be tested by studying existing psycholinguistic literature

about comprehension under stress.

1.2.2. Thesis goals

The aims specified in Section 1.2.1 will be achieved by setting and meeting the following goals:

Goal 1 is to identify and select a set of text factors which contribute to high complexity of Crisis Management documents, their low comprehensibility by target readers, and poor performance on secondary tasks. This will involve examining the existing sets of text complexity features affecting both human readers and NLP applications.

Goal 2 is to perform a critical review of the existing approaches in text complexity and text simplification and to investigate their applicability and/or their limitations with respect to documents in the Crisis Management domain.

Goal 3 is to collect data needed for the analysis of written documents in the Crisis Management domain. For this reason, an analysis of the existing types of crisis management documents will be performed, and a representative corpus of CM documents will be collected.

Goal 4 is to investigate the extent of high text complexity (TC) factors in Crisis Management documents. For this reason, a set of automatic approaches to recognize and count the high issues in emergency texts will be designed and implemented.

Goal 5 is to propose and develop an appropriate approach for writing and simplifying texts, based on linguistic theory, which must be tailored to crisis management documents written in English. The

approach must be able to address the majority of the linguistic complexity issues identified by Goal 4.

Goal 6 is to perform an evaluation of the proposed approach for writing and simplifying texts in terms of whether it has a positive impact on the comprehensibility of emergency instructions. Appropriate evaluation techniques will be identified, adapted, or developed, and experiments will be run.

Goal 7 is to evaluate the impact of the applied approach for writing and simplifying texts on other tasks important for the domain, such as manual and automatic translation to other languages. Appropriate evaluation methods will be identified or developed, and experiments will be designed and conducted.

Goal 8 is to evaluate the acceptability of the proposed approach for writing and simplifying texts with end-users and to identify its concrete weaknesses in terms of applicability. For this reason, an especially tailored experiment will be designed, and end-users with appropriate qualifications will be recruited and trained. Also, materials for this experiment will be produced.

Goal 9 is to identify any weaknesses and limitations of the methodologies proposed in Goals 4-8 and to identify directions for improvement and future research.

1.2.3. Contributions of the thesis

By achieving the goals set in Section 1.2.2, the thesis makes the following novel contributions to knowledge:

The **first main original contribution** of this thesis is the first scientific investigation of the phenomena of text complexity affecting English documents from the crisis management domain. The investigation includes identifying a set of linguistic features which may appear in CM documents and which can affect both human comprehension and the performance of NLP applications.

The **second main original contribution** of this thesis is the development of writing guidelines for re-writing existing or producing new clear crisis management documents in English. The importance of this approach is that it will be tailored to the situational circumstances and reading and linguistic characteristics of the texts in the domain.

The **third original contribution** of this thesis is the development and deployment of evaluation techniques for testing methods for re-writing documents into clear ones or for producing clear documents in the crisis management and other domains. This contribution is important for the currently developing field of automatic text simplification and natural language generation for lay readers.

The **fourth original contribution** (and a by-product) of this thesis is the development of linguistic resources tailored for the domain, such as:

- a set of domain-specific hight text complexity features,
- a corpus of representative types of crisis management documents in English,
- training and testing materials for end-users, and
- sets of original and simplified versions of crisis management documents.

The development of these resources is crucial for the domain, as they can be used by the research community for developing and testing other NLP applications for English documents in the crisis management domain. They will also be important for other NLP applications, such as automatic text simplification.

1.3. Structure of the thesis

This thesis consists of four parts. The **first part** consists of Chapter 2. It presents a critical overview of related work in text complexity and text simplification, with a focus on controlled languages and the crisis management field. The **second part** comprises Chapter 3. It introduces the first novel contribution of the thesis, namely the investigation of the text complexity of representative types of crisis management documents. Chapter 4 represents the **third part** of the thesis. It describes the proposed writing guidelines for simplifying and producing of clear crisis management documents. Chapters 5, 6, and 7 constitute the **fourth part** of the thesis. They offer extensive evaluation of the impact of texts, produced according to the proposed writing guidelines, on the comprehensibility of emergency instructions, on the tasks of manual and automatic translation, and on acceptability to users. The contents of each chapter are analysed more concretely below.

This chapter (**Chapter 1**) introduces the context and motivations for this study and the assumptions behind this research. It also lists and describes the aims and goals to be achieved in this thesis and its original contributions, and finishes by presenting the structure of the thesis.

Chapter 2 defines the concepts of *Text Complexity (TC)* and *Text Simplification (TS)*, which are respectively the problem addressed by this thesis and the solution to it which has been proposed by this thesis. The chapter presents and analyses the sets of text complexity features which hinder

human comprehension (with which **Goal 1** is partially accomplished) and the performance of computer applications and presents a critical overview of approaches for measurement and reduction of text complexity (i.e. text simplification applications). The limitations of the existing approaches and their applicability to measuring and reducing text complexity in documents in the crisis management domain are examined. With this, **Goal 2** is achieved. The approaches for reducing text complexity are divided into manual, semi-automatic, and fully-automatic ones, and are classified into two groups on the basis of whether they are based on controlled language or not. The chapter concludes that the best candidate to be applied for the purposes of the thesis is a controlled language for French.

Chapter 3 introduces the investigation of the text complexity of crisis management documents. The chapter describes the methods of text collection, pre-processing, and composition of the Crisis Management Corpus, containing representative types of crisis management documents (**Goal 3**). The chapter describes the methods followed and Python scripts developed for performing the text complexity analysis (**Goal 4**), and finalises the set of high text complexity features selected to be investigated from among those presented in Chapter 1 (which completes **Goal 1**).

Chapter 4 presents the proposed controlled language writing guidelines for crisis management documents in English, CLCM. By this, **Goal 5** is achieved. The chapter presents the project in the context of which the approach was developed, as well as the high text complexity issues which it addresses, and how it addresses them. The differences between the proposed approach and the controlled language from which it was adapted and other similar approaches are outlined. The chapter also presents the results of a study aiming to transfer this approach to an under-resourced European language (Bulgarian).

Chapter 5 is the first of the three evaluation chapters. It evaluates the impact of texts, produced according to the writing guidelines proposed in Chapter 4, on the most important task for its purposes—reading comprehension under stress (**Goal 6**). Due to the specificity of the task, it is argued that the existing controlled language evaluation approaches are not appropriate, and a special experiment ("*Online Reading Comprehension Experiment*") is run. The setting of the experiment is described, and the results and findings are reported.

Chapter 6 is the second of the evaluation chapters. It describes the evaluation of the controlled language proposed in Chapter 4 on two tasks which are important for the domain: manual translation and machine translation. The motivation for selecting these two tasks is that in the modern global world, crisis management documents often need to be translated in order to reach a larger audience. It is thus important that the controlled language rewriting improves and does not hinder the performance of these tasks. The chapter discusses the existing evaluation approaches and proposes and adapts and develops new ones appropriate for the domain. The chapter describes an evaluation experiment (the *"Translation and Post-editing Experiment"*) and reports its results and findings. With this, **Goal 7** is achieved.

Chapter 7 is the last of the evaluation chapters and also the final one of the chapters which contains original contributions of this thesis. The aim of this chapter is to evaluate the guidelines proposed in Chapter 4 for writing and re-writing text on end-users' acceptability, and the difficulty of applying this approach for simplifying texts. Due to the lack of existing approaches, a tailored evaluation approach was developed. The evaluation is based on examining simplified versions produced during a specific experiment (the *"Text Simplification Task Experiment"*) by several linguists and on the basis of their acceptability judgements as elicited by a questionnaire. Results and findings regarding the concrete difficulties encountered by users and directions for future implementation

priorities are reported. This chapter fulfils **Goal 8**.

Chapter 8 is the last chapter of the thesis. It reviews the extent to which the goals set in Chapter 1 are met and provides details about the original final contributions of this research, as well as a review of the thesis. It also provides directions for future work on the basis of the weaknesses discovered while conducting the research described in the previous chapters. By this, it achieves **Goal 9.**

Chapter 2 – Text Complexity and Text Simplification

If I have seen any further it is only by standing on the shoulders of giants. (Isaac Newton)

The aim of this chapter is to introduce the problem addressed by this thesis (high text complexity), to propose a solution for reducing it (text simplification) and to examine the existing approaches to measuring and reducing text complexity, as well as their limitations and applicability with respect to the domain of study of this thesis. The chapter starts by introducing the concept of Text Complexity (TC) and the factors which affect human readers and computer applications (Section 2.1). Next, Section 2.2 presents various approaches to measuring text complexity. Section 3 introduces the existing approaches to reducing text complexity, i.e. Text Simplification (TS) approaches. Finally, Section 4 presents the summary and the conclusions of the thesis, providing the justifications for the choice of the approaches which are applied for the purposes of this thesis.

2.1. Text Complexity and Factors which Affect Text Complexity

This thesis defines Text Complexity (TC) (or "Text Difficulty", G. Leroy et al., 2010) the internal characteristic of a written text which affects human comprehension during reading or the performance of computer applications processing text. Text complexity is independent of the typographic presentation of the text (font style and size, spacing, indentation, and others), the environmental conditions (e.g. light), and the reading skills of the reader. Text complexity can manifest itself on all three text levels: lexical (affecting words' meaning), syntactic (affecting sentence structure), and discourse (affecting the text structure and cohesion as a whole). There are differences and similarities between the types of text complexity which can pose difficulties to human readers and those which can pose difficulties to computer applications processing text. This research aims to examine both types of TC issues, because its goal is to address reducing TC for both targets. The following two sections describe the main text complexity issues which human readers and computer applications may encounter. Section 2.1.1 discusses text complexity issues for humans, while Section 2.1.2 gives an overview of the text complexity issues for computer applications. Further on, Section 2.1.3 provides a more detailed overview of the individual textual and linguistic characteristics which contribute to increasing text complexity.

2.1.1. Text complexity issues for human readers

In order to discuss the TC issues relevant for humans, it is necessary to first describe how reading works. Figure 2.1 exemplifies the process of reading.

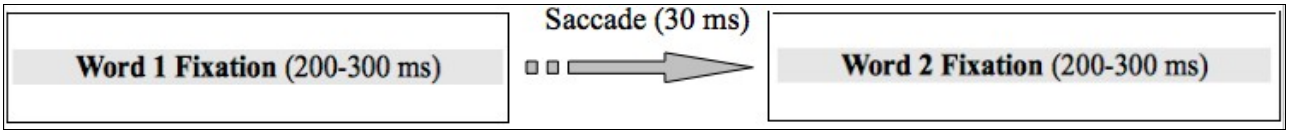


Figure 2.1: The process of reading. Fixations and Saccades.

As shows Figure 2.1, according to research studies in the field of psycholinguistics (Rayner, 1998), while reading, the eyes perceive and process incoming information only when they stay still, fixed on one point in the text. The eyes stay still for about 200-300 ms. These periods are called "fixations". The eyes also perform jumps from one point of the text to another, during which the eyes move so fast that it is impossible to process information. These jumps are called "saccades" and have a duration of only around 30 ms. (Rayner, 1998). The information which can be processed during fixations is also limited. The eye can catch up to 15 characters to the right and 3-4 to the left (Harley, 2008). Within these fixations the visual field has different capabilities and acuity of vision in different regions. From the three visual area regions (fovea, parafovea and periphery), the characters are being recognized most clearly in the fovea region and less clearly in the parafovea area. In the fovea region the central seven characters are processed. It has been proven that more experienced readers should be able to process a larger span of characters than poorer readers (Martin, 2004). There are different models of reading. The most commonly accepted one is the E-Z Reader model (Reichle et al., 2003), in which when a person reads, the eyes are first fixated over the first point in the text. The visual attention progresses forward until the moment in which the acuity limitations of the visual system do not allow recognizing words and processing visual information. Eyes then shift to that point and attention proceeds from that point on. When something is not clear in the text, it is necessary to move the eyes back to a previous point of the text. This movement is called "regression" and is exemplified in Figure 2.2. Explaining how the process of reading works helps to understand what happens if a comprehension difficulty is encountered in text: 1. The fixation takes longer time, and 2. The reader is forced to move his eyes back to the previous fixation points, in order to disambiguate or better comprehend the difficult

point.

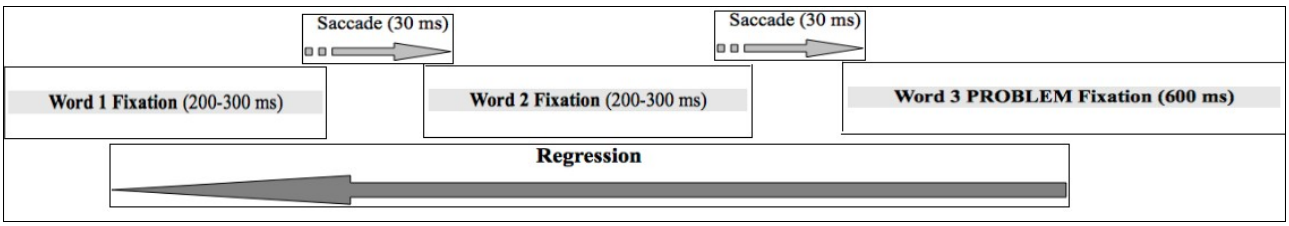


Figure 2.2: The process of reading. Regression.

Reading starts with word recognition (as shown in Figures 2.1 and 2.2) then proceeds with phrase and sentence comprehension and ends up with discourse and text comprehension. According to Harley (2008) word recognition does not only consist of recognizing that a given string of letters is familiar or not, but also consists of relating all linguistic information about this word with the current string. Such information includes: the meaning of the word, the part-of-speech of the word (which allows the reader to deduce in which roles this word can appear in a sentence) and how the word is pronounced. For example, when meeting the word “sentence”, the reader recognizes that it is composed of the symbols “s e n t e n c e”, that it means a string of words, that it is a noun and which it is pronounced /ˈsɛntəns/. Psycholinguists (Johnson-Laird, 1975; Gerrig, 1986) say that it is not the case that all of the information about a word is always assessed at the same time. More concretely, word recognition is composed by visually accessing a word's familiarity, assessing its representation in the mental lexicon, accessing the word's meaning, (mentally) pronouncing the word, and, on the basis of assessing its pronunciation, assessing its meaning.

There are some phenomena which can make word recognition easier or more difficult. Processes which help word recognition are: facilitation by words in the immediate context which have related meanings (called semantic priming), syntactic facilitation (the syntactic structure of the sentence provides hints about the part-of-speech of the expected word), high frequency, and early age-of-acquisition. High-frequency words are much faster and easier to recognize; this has been proven by many experiments (Howes & Solomon, 1951; Whaley, 1978; Forster & Chambers, 1973). Age-of-

acquisition is the age at which a person first learns a word. Many experiments have shown that words with early age-of-acquisition are recognized faster than those with which are learned at a later stage (Barry, Morrison & Ellis, 1997; Brown & Watson, 1987; Carroll & White, 1973). Some issues can hinder word recognition for particular types of readers, but facilitate word recognition for most experienced readers. Such a case are words with large orthographic neighbourhood (words which can be created by changing one letter, e.g. *pine/mine/dine/line/bine/kine/fine/nine/tine/sine*), which can produce difficulty at the level of visual recognition of the string of letters for readers suffering from surface dyslexias, as they can commit visual errors and recognize a word wrongly (Harley, 2008). In contrast, experiments with experienced readers have shown that low frequency words which have frequent orthographic neighbours are recognized faster than other words (Andrews, 1997). Issues hindering word recognition for even experienced readers are words which may belong to more than one part-of-speech category (e.g. *a fire* (noun)/*to fire* (verb); *a can*/*I can*) and words which can have more than one semantic meaning (MacKay, 1966), whether they are semantically unrelated (e.g. *bank* = 1. financial institution; 2. the border of a river) or related (e.g. *film* = 1. a movie 2. the material on which photographs are stored). The last two causes of difficulty (POS or semantic ambiguity) make necessary the process of selecting the right meaning of ambiguous words.

For the reasons described above, difficulties with text comprehension for human readers arise firstly at the lexical level: problematic issues are usually non-familiar words, such as infrequent words, technical terminology and abstract concepts. Everyday words and concrete concepts are usually easier to understand, as can be seen in the following examples of sentences (*1a* and *1b*), taken from the Plain English Campaign website. They illustrate the difficulty that uncommon words may pose and how much easier the comprehension of common words is. The problematic words are provided in bold:

1a. ***High-quality learning environments*** *are a* ***necessary precondition*** *for* ***facilitation*** *and* ***enhancement*** *of the* ***on-going learning process****.*

1b.Children need good schools if they are to learn properly.

Sentences 1a and 1b have the same meaning, with 1a being far more difficult to understand than 1b, because the same meaning has been expressed using infrequent terms.

At the syntactic level the most frequent complexity issues are long sentences, convoluted syntax, and high number of modifiers. They can cause processing difficulties, because they are known to overload working memory (Siddharthan, 2003, Harley, 2008). Working memory overload means that too many information units (concepts, clauses, relationships) have to be kept in mind while processing a sentence. Although it is known that the longest existing sentences reach several thousand words (Weisler et al., 2000), obviously such examples cannot be listed because of lack of space. The Plain English Campaign website provides an example of a long sentence, composed of 630 words. Part of it is displayed in Figure 2.3.

A path from a point approximately 330 metres east of the most south westerly corner of 17 Batherton Close, Widnes and approximately 208 metres east-south-east of the most southerly corner of Unit 3 Foundry Industrial Estate, Victoria Street, Widnes, proceeding in a generally east-north-easterly direction for approximately 28 metres to a point approximately 202 metres east-south-east of the most south-easterly corner of Unit 4 Foundry Industrial Estate, Victoria Street, and approximately 347 metres east of the most south-easterly corner of 17 Batherton Close, then proceeding in a generally northerly direction for approximately 21 metres to a point approximately 210 metres east of the most south-easterly corner of Unit 5 Foundry Industrial Estate, Victoria Street, and approximately 202 metres east-south-east of the most north-easterly corner of Unit 4 Foundry Industrial Estate, Victoria Street, then proceeding in a generally east-north-east direction for approximately 64 metres to a point approximately 282 metres east-south-east of the most easterly corner of Unit 2 Foundry Industrial Estate, Victoria Street, Widnes and approximately 259 metres east of the most southerly corner of Unit 4 Foundry Industrial Estate, Victoria Street, then proceeding in a generally east-north-east direction for approximately 350 metres to a point approximately 3 metres west-north-west of the most north westerly corner of the boundary fence of the scrap metal yard on the south side of Cornubia Road, Widnes, and approximately 47 metres west-south-west of the stub end of Cornubia Road be diverted to a 3 metre wide path from a point approximately 183 metres east-south-east of the most easterly corner of Unit 5 Foundry Industrial Estate, Victoria Street and approximately 272 metres east of the most north-easterly corner of 26 Ann Street West, Widnes, then proceeding in a generally north easterly direction for approximately 58 metres to a point

Figure 2.3: Example of a long sentence.

The source of the sentence is a legal contract which has been awarded one of the annual Plain English Campaign's awards for “worst written document”. As can be seen, it is characterized both by working memory overload in terms of concepts and by a complexity of syntactic relations.

Another kind of problem which humans encounter is ambiguous expressions and constructions. It is known that ambiguities take more time to be processed (Harley, 2008), as readers need more time to check more than one of the existing alternatives. Ambiguities can arise at both lexical and syntactic levels. An example of a lexical ambiguity has been employed by MacKay (1966), who has carried out an experiment involving measuring time human participants employed to complete the following two sentences:

*2a. After taking the **right** turn at the intersection, I ...*

*2b.After taking the **left** turn at the intersection, I ...*

The results showed that the participants took longer to complete 2a than 2b, which was explained by the ambiguity of the word *right*.

Syntactic ambiguities can be of different types, examples being bracketing and parsing ambiguities (Harley, 2008). An example of a bracketing ambiguity is given in the sentence *Old men and women leave first*. The ambiguity consists of whether the adjective *old* modifies *men* alone or also *women*. In this case two interpretations are possible: *Old [men and women] leave first.* and *[Old men] and women leave first.* An example of parsing ambiguity is manifested in the following newspaper article title (Harley, 2008): *Police seek orange attackers*. The sentence can be interpreted in several ways, in accordance with the different interpretation of the roles and relations between the word "*orange*" and the word "*attackers*":

- The police seek attackers who are orange
- *The police seek attackers who attacked an orange,* and
- The police seek attackers who attacked with an orange

Another good example of a parsing difficulty is *The old man the boats*. In this sentence, the reader can be confused by the fact that the two words "*old*" and "*man*" are syntactically ambiguous and can be parsed in two different ways. "*Old*" can be either an adjective (meaning "*aged*" or "*senior*"), or a noun (meaning "*old people*"). "*Man*" could be either a noun (meaning "*male human*"), or a verb (meaning "*operate*"). This kind of ambiguity is known as "garden path ambiguity", because the reader is misled by the semantic and syntactic context until she/he reaches the end of the

sentence (Harley, 2008).

Another important type of ambiguity is PP (prepositional phrase) attachment ambiguity. A classic example is “I saw the man with the telescope”. In this sentence, “with the telescope” can mean 1. that the man had a telescope, in which case the PP is attached to the noun phrase, or 2. that the man was seen through a telescope, in which case the PP is attached to the verb phrase.

Humans resolve ambiguities by a look-up at the local and global contexts. Local context consists of the meaning and syntactic roles of the previous words, while the global one corresponds to the world knowledge. It is also important to note that some target groups of readers, such as low-skilled readers, children, non-native speakers or people suffering from language disorders (such as aphasia and dyslexia) encounter more difficulties than a high-skilled reader usually does. Some issues are more problematic for particular groups of low-skilled readers. For example, patients suffering from aphasia find difficulties resolving pronouns, readers suffering from dyslexia have more problems with long words, and the words which a native speaker may find common may be considered uncommon for non-native speakers.

2.1.2. Text complexity issues for NLP applications

Natural Language Processing (NLP) is an interdisciplinary field whose aim is to develop computer applications which can process human language. There are various NLP sub-areas of research: Machine Translation (MT) (applications which translate a given text from one language to another), Text Summarisation (automatic generation of summaries), Information Extraction (automatic extraction of structured information from text), Speech Recognition, Speech Synthesis, and many others (Jurafsky and Martin, 2008).

Regardless of the purpose of the application, it is quite common that the text needs to be preprocessed and includes many of the following tasks: identification of words in the string of symbols (*text tokenisation),* splitting the text into meaningful units (*text segmentation),* assignment of part of speech tags to every word (*part-of-speech tagging*), assignment of the syntactic role of each word and the syntactic relationships between them (*parsing),* assignment of semantic roles *(semantic role labelling),* mapping of words into meanings (*word sense disambiguation),* identification of missing syntactic elements (*ellipsis resolution*), and linking pronouns or other kinds of anaphora to their antecedents (*anaphora resolution*).

In processing text, NLP applications are more restricted than humans, since they can rely only on pre-defined procedures and limited resources and make decisions based mainly on the local context (Jurafsky and Martin, 2008). For this reason, text complexity issues which cause processing difficulties for NLP applications are reduced primarily to resolving ambiguities. A very good example of a highly ambiguous sentence which could cause processing difficulties to different kinds of NLP applications is provided by Jurafsky and Martin (2008): *I made her duck*. Different NLP applications encounter ambiguities at different levels of processing. For example, a speech recognition system has to be able to recognize that the word /meɪd/ in the sentence /aɪ meɪd hər dʌk/ is *made* and not *maid*. *Her* is morphologically ambiguous, since it is the same form for a dative pronoun or a possessive pronoun. The word *duck* creates ambiguity at the syntactic level, because it can be a noun or a verb. The verb *make* is semantically ambiguous – it can mean "*create*" or "*cook*", but it is ambiguous from the syntactic point of view, since it can be transitive (requiring only a direct object) or ditransitive (requiring two objects: a direct and an indirect one). In this way, the above mentioned sentence can be interpreted in several different ways: "*I cooked a duck for her*." or "*I cooked a duck, which was belonging to her.*" or "*I made her a (plasticine) duck*." or "*I made her bend suddenly.*" and even "I turned her into a duck.". The example *"I saw the man with the*

telescope.", discussed in the previous section is also a problem for NLP applications.

For this reason, the presence of more modifiers (adjectives, prepositional phrases, relative clauses or subordinate clauses) causes high sentence length and thus cane create more parsing candidates. It has been shown, for example, that the performance of machine translation systems is decreased for longer sentences (Gerber and Hovy, 1998). The text complexity factors negatively affecting the performance of machine translation engines can be grouped into the concept of Mtranslatability (Bernth and Gdaniec, 2001) and are MT-engine dependent.

The next section will examine concrete text complexity issues: an overview of ways to measure text complexity, based on the presence in text of specific complexity markers, will be provided in Section 2.3, while the approaches to resolution of different text complexity issues will be discussed in Section 2.4 in relation to concrete NLP applications, such as text reduction for small screens, text summarisation, and others.

2.1.3. Factors which affect text complexity

This section will analyse the concrete high text complexity issues. Section 2.1.3.1 will make a general comparison between high TC issues affecting human readers and high TC affecting NLP applications, while Sections 2.1.3.2, 2.1.3.3 and 2.1.3.4 will discuss separately the lexical, syntactic and discourse TC issues.

2.1.3.1. General analysis

In order to analyse the concrete high text complexity issues, a tentative comparison of the text complexity issues specific for both humans and NLP applications and the approaches to solve them is provided in Table 2.1, followed by a more detailed discussion of these individual issues. The first column provides the type of the text complexity issue (lexical, syntactic or discourse). The second column specifies the individual text complexity issue, while the third and the fourth ones contain markers indicating whether the specific type of text complexity issue can be considered a difficulty for humans and NLP applications. "*YES*" means that this issue constitutes a difficulty, while "*NO*" that this text complexity issue is irrelevant for either humans or NLP applications.

Type of linguistic complexity	**Issues**	**Human readers**	**NLP Applications**	**References**
Lexical	Rich vocabulary	YES	YES	Tweedie and Baayen (1998)
Lexical	Long words	YES	NO	Harley (2008), Flesch (1948), Kincaid et al. (1975)
Lexical	Infrequent, technical terms	YES	YES	Devlin (1999)
Lexical	Ambiguous words	YES	YES	MacKay (1966), Harley (2008)
Lexical	Vague quantifiers	YES	YES	Graesser (2006), Cramer (2009)
Lexical	Words with high age-of-acquisition	YES	NO	Coltheart (1981), Harley (2008)
Lexical	Abstract concepts	YES	YES	Paivio (1971), James (1975)
Lexical	Words with large orthographic neighbourhood size	YES	NO	Harley (2008)
Lexical	Inconsistent terminology	YES	YES	Renahy et al. (2011)
Lexical	Figurative language	YES	YES	Harley (2008), Dobrovol'skij et al. (2005), Lönneker-Rodman et al. (2008)
Syntactic	Long sentences	YES	YES	Harley (2008), Jurafsky and Martin (2008)
Syntactic	Complicated syntax	YES	YES	Harley (2008), Jurafsky and Martin (2008)
Syntactic	Too much information to remember	YES	NO	Harley (2008)
Syntactic	Passive voice	YES	YES	Quirk (1985), Harley (2008), Cohen et al. (2010)

Syntactic	Negative constructions	YES	YES	Glenberg (1999), Szarvas (2008)
Discourse	Pronouns with unclear reference	YES	YES	Quirk (1985), Mitkov (2002), Canning (2002)
Discourse	Illogical order	YES	YES	Heurley (2001)
Discourse	Missing discourse connectives	YES	YES	Schiffrin (1987), Burstein et al. (2010)

Table 2.1: Comparison of text complexity issues for human readers and NLP applications

As can be seen, most of the complexity issues are a problem both for human readers and for language processing applications. Generally the perception of which text is complex depends on the type of reader, and thus not all of the text complexity issues represent reading difficulties for all readers. The same is true for the complexity issues for NLP applications. More detailed analysis of the individual text complexity issues follows below. The TC issues are discussed as presented in Table 2.1 and thus divided into lexical (Section 2.1.3.1), syntactic (Section 2.1.3.2) and discourse (Section 2.1.3.3).

2.1.3.2. Lexical high text complexity issues

As explained in Table 2.1, the high lexical TC issues are nine. A discussion of each follows.

Rich vocabulary

Rich vocabulary can be measured via the vocabulary size. This thesis defines **vocabulary size** (also called **lexical richness** or **lexical diversity**) as the number of different words in a given text. If a text has a large vocabulary size or a rich vocabulary, this may imply using different synonyms with subtle meaning differences between them in the same situation. The example below shows different expressions used to indicate the same concepts found in emergency instructions. The alternative expressions of the same concept are listed on the same row.

- building/place/location
- place/home/your house
- vehicle/car
- go off/explode
- phone numbers/telephone numbers
- patient/person/someone
- threat/danger
- way out/escape route

It is observed that rich vocabulary can cause comprehension problems for human readers for the following reasons:

1. If they are non-native speakers or non-specialists in the domain, they may be not aware of the meaning of rarer synonyms and their relationships with the main term in the synonym set.

2. If two different synonyms are used in the same context to denote the same situation, the readers may think that they denote two different situations.

It is observed that rich vocabulary can be a problem for language processing applications for the following reasons:

1. If not all the different synonyms are in the dictionary (if any dictionaries are used), the program may not recognize them.

2. If the language processing application has no information about the synonymic relationships between the terms employed in the text and those in the dictionary, this may inhibit the recognition of the fact that several terms point to the same entity in the real world.

The most common way (Tweedie and Baayen, 1998) to measure lexical richness or vocabulary size of a text is to compute the number of different word types divided by the total number of words occurring in the text.

Long words

This thesis defines **long words** as word with two or more syllables. Long words are measured via word length as the number of symbols of which a word is composed. It is considered that a high number of symbols or of syllables, resulting in morphologically complex and thus long words, can hinder reading and could create comprehension problems (Harley, 2008). One of the reasons for this is that although there are some very commonly used long words, such as "*television*", most long words are actually technical terms and can be not known by all readers. Examples are the words: "*antidisestablishmentarianism*" (opposition to the disestablishment of the Church of England, 28 letters), "*floccinaucinihilipilification*" (the estimation of something as worthless, 29 letters), "*pneumonoultramicroscopicsilicovolcanoconiosis*" (a lung disease, 45 letters)[4].

It is generally considered that words with two or more syllables can be a problem for particular groups of readers, such as less experienced readers or people with dyslexia (Harley, 2008). The

4 Source: ("What is the longest English word?". AskOxford. http://www.askoxford.com/asktheexperts/faq/aboutwords/longestword. Last accessed on 2011-03-22).

word's length usually does not represent a problem for language processing applications, except perhaps in the case of applications processing words or making predictions at the morphological level.

Word length is usually measured in one of two ways: either by computing the number of letters a word is composed of, or by computing the number of syllables in a word (Flesch, 1948, Kincaid et al., 1975). Word length is one of the fundamental TC issues to be calculated in the readability formulae (see Section 2.2).

Infrequent, technical terms

This thesis defines **infrequent terms** as those which have the lowest frequencies in the everyday language. **Technical words** are an example of infrequent terms and can be defined as domain-specific words used by specialists in recurrent situations.

Both infrequent and technical words can constitute a burden both for human readers (Devlin, 1999) and for NLP applications, since humans may not know their meaning, while NLP applications either may not have them in their dictionaries nor may not be able to predict their behaviour because of their low frequency. To measure the amount of their presence in text necessitates either specific lists of words or frequency lists from which only the words with the lowest frequency are drawn. Technical and infrequent words (or rather their most common and well-known counterparts) are one of the fundamental TC issues to be calculated in the readability formulae (see Section 2.2).

Ambiguous words

This thesis defines as **ambiguous words** those words which have more than one sense or meaning.

Ambiguous words can be a problem for both human readers and computer applications. Ambiguous words constitute a problem for human readers because, in order to process them, the reader needs to make regressions of her/his eyes to previous points in the text, which increases reading time and thus the difficulty of reading. In the psycholinguistic literature it is now known that when an ambiguous word is encountered, all senses are activated, and then the context is used for disambiguation. (Harley, 2008). The examples *2a* and *2b* (MacKay, 1966) in Section 2.1.1 have shown that a sentence containing an ambiguous word takes a longer time to process than a sentence containing no ambiguous words.

Problems which computer applications can have with ambiguous words can be seen in one of the experiments with a machine translation engine presented in Chapter 6 and in (Temnikova and Orasan, 2009). The text re-written in the controlled language has less context than the original one, and for this reason, while translating from English to Russian, the machine translation engine Google Translate made a mistake - "*Stay inside*" was wrongly translated as "Пребывание внутри." (= "*Staying inside*") because of the part of speech ambiguity of the word "stay", which can be both a noun and an imperative form of the verb "to stay". As can be seen, ambiguous words constitute a major difficulty for computer applications, especially when there is not enough context. A way to measure the amount of ambiguous words in text may be to use WordNet (Fellbaum, 1998) to identify them in text and then to count their frequencies.

Vague quantifiers

The **vague quantifiers** overlap with a subset of the indefinite pronouns (some, a few, anybody), and are defined by Graesser (2006) as quantitative adjectives or adverbs, whose numerical value is not

specified on a functional scale or underlying quantitative continuum. The vague quantifiers represent a problem both for humans (Graesser, 2006) and computer applications (Cramer et al., 2009). As it is outside the topic of this thesis to discuss the vague quantifiers in great detail, the way to measure their quantity in the text here is also to have a pre-compiled list of them and calculate their frequency of occurrence.

Words with high age-of-acquisition

Words with high age-of-acquisition can be defined as those which people learn at a more mature age and not in childhood. It is considered that words which are learned later in human development are more difficult than those which are learned first (Harley, 2008). A way to measure their presence in a text is to use the MRC Psycholinguistic Database (Coltheart, 1981), which contains a list of them.

Abstract concepts

This thesis defines **abstract concepts** as those which do not have a referent in the real world, in contrast with concrete concepts denoting physical entities. An example of a concrete concept is the word "*dog*" (there is an object corresponding to it in the real world), while an example of an abstract one is the word "*love*" (which does not have an object corresponding to it). Abstract concepts represent a larger problem for humans than for computer applications. This is caused by the fact that there are almost no computer applications processing abstract concepts in a way which accesses their non-concrete meanings. The only such applications are those which aim to interpret figurative language (Nayak, 2011), but there is only limited activity in this area of research. In contrast, in psycholinguistics it has long been known that abstract concepts are more cognitively

difficult to process than concrete ones (Paivio, 1971; James, 1975). They constitute a processing problem for less-literate readers, non-native speakers, or readers with certain disorders (like aphasia). A way to measure their presence in text would necessitate first the ability to identify them in prose. As this is difficult to determine programmatically, their presence will be considered to be difficult to quantify. A very simple approach would be to build a list of abstract terms from a dictionary and count the occurrences.

Words with large orthographic neighbourhood size

Orthographic neighbourhood is defined as the set of words which differ only by one letter from a specific word (Harley, 2008). Words with large orthographic neighbourhood size constitute a problem only for human readers. Some types of dyslexic readers in particular have difficulties processing orthographically similar words (Harley, 2008). For this reason, the larger the orthographic neighbourhood size of a certain word is, more difficulties they experience. An example of a word with a large orthographic neighbourhood is "*mine*", which has an orthographic neighbourhood of 29: "*line, pine, mile*", etc. A way to measure the quantity of such words in text could be to write a regular-expressions-based or finite state automata algorithm to recognize all the possible candidates in a very large corpus.

Inconsistent terminology

Inconsistent terminology can be defined as the phenomenon of using synonymic expressions to denote the same concept and it is related to the problem of **rich vocabulary**, which has already been discussed above. An example could be: "*I am going home."/"I am going to my house.*" where "home" and "house" are synonyms, denoting the same concept, but they can also denote slightly

different entities, as synonyms usually do (e.g. the speaker may have both an apartment where he/she lives, i.e. which is his or her home, and own a house, in which he/she does not live). Synonyms which are used to denote the same concept can create problems of ambiguity and the inability to recognize that the statement refers to the same situation, especially for non-specialist readers (Renahy et al., 2011). For this reason, inconsistent terminology can cause problems to both human readers in the case of technical text and to computer applications which need to extract all occurrences of a given event. It is difficult to quantify the presence of these issues in a text. A direct way to measure it could be to find all possible synonyms occurring in the same close context, which is a relatively hard NLP task, which requires very large corpora and gives only approximate results (Banko and Brill, 2001); an easier but indirect way to measure it would be to calculate the vocabulary size (explained earlier).

Figurative language

Figurative language is defined as use of language which goes beyond the literal meanings of the words (Harley, 2008). According to Lönneker-Rodman (2007), the main types of figurative language are metaphor, metonymy, idioms, sarcasm, and humour. Examples of figurative language are the following sentences:

- Sentence A: "*That flat tire cost me an hour.*" (metaphor)
- Sentence B: "*She is reading Shakespeare.*" (metonymy)
- Sentence C: "t*o spill the beans*" (idiom)
- Sentence D: "*The carbon duck was delicious.*" (sarcasm)

Figurative language can be a processing issue for both human readers and computer applications (Harley, 2008; Dobrovol'skij et al., 2005; Lönneker-Rodman et al., 2008). Figurative linguistic expressions could be a problem for human readers because they first need to access their literal meaning, then test it against context, and finally look for an alternative meaning, which makes the cognitive processing longer and more difficult (Harley, 2008). In fact, some categories of readers, for example autistic ones, are not able to understand the abstract meaning of a figurative statement. Computer applications may also have problems with that, as they may not have access to the figurative meaning and may recognize only the literal one. This lexical issue is related to the one of "abstract concepts". The incidence of figurative language is currently very difficult to quantify, as only domain-dependent lists of frequently used expressions can be used for calculating their frequencies in the text, and no approach for the moment can identify newly-formed or rarer figurative language expressions.

2.1.3.3. Syntactic high text complexity issues

As said in Table 2.1, the high syntactic TC issues are six. A discussion of each follows.

Long sentences

This thesis defines **long sentences** as sentences containing too many words. The number of words of a short or a long sentence as well as the sentence length distributions depend on the text's domain. For the purposes of this thesis, sentences containing one to five words are defined as very short ones, those containing six to ten words as standard ones, and those containing more than twenty words as overly long sentences[5]. Long sentences can constitute a problem both for human readers and for computer applications (Harley, 2008; Jurafsky and Martin, 2008). Long sentences

5 http://www.plainenglish.co.uk/, last accessed on November 13th, 2011.

represent a comprehension burden for readers because they offer too much information with respect to number of words, relationships among them, logical order of elements, etc. to keep in short-term memory during cognitive processing. For computer applications, long sentences usually create more ambiguities at the syntactic level and lead to errors in the output of parsers.

An example of an averagely long sentence taken from the Microsoft Excel manual follows below:

"*When you use open a workbook which was created in an earlier version of Excel, all of the formulas in the workbook — those that depend on cells that have changed and those that do not — are recalculated.*"[6]

The obvious approach to calculating the length of a sentence is to calculate the number of the words of which it is composed (Flesch, 1948, Kincaid et al.,]1975). Sentence length is one of the fundamental TC issues to be calculated in the readability formulae (see Section 2.2).

Complicated syntax

This thesis defines **complicated syntax** as the phenomenon which occurs when the syntactic structure of a sentence is not linear, but convoluted and evolving on several different levels, and the relationships between the different entities participating in the sentence are too numerous, which can make them unclear.

Complicated syntax can constitute a difficulty both for human readers and for computer applications, as it is frequently difficult to disambiguate the syntactic relationships between the elements in such sentences. In fact, it is known that syntactic analysis is necessary before being able

6 Taken from http://office.microsoft.com/en-us/excel-help/change-formula-recalculation-iteration-or-precision-HP010054149.aspx .Last accessed on November 13th, 2011.

to determine the thematic roles of the entities in a sentence (Harley, 2008), which means that having difficulties at that level can delay the processing at the next level. Human readers also have difficulties because they have to remember too many information units, such as the single words and the relationships between them, before being able to understand the meaning of the whole sentence. Complicated syntax can cause a burden for computer applications, as well, because the more complicated the syntactic structure of a sentence is, the more ambiguities can arise at the syntactic level. According to Szmrecsanyi (2004), there are three ways to calculate the degree of syntactic complexity of a sentence:

- the number of words in the sentence,
- the number of nodes in the syntactic tree,
- through the Index of Syntactic Complexity (ISC), which takes into consideration the number of nouns, verbs, subordinating conjunctions, and pronouns.

However, it has been demonstrated that these three measures are more or less equivalent and well correlated (Szmrecsanyi, 2004; Cohen et al., 2010) the obvious solution would be to take the number of words in the sentence as a measure of both sentence length (or “long sentences”, as defined in the previous TC item) and syntactic complexity of a sentence. Additional TC markers which contribute to estimating the complexity of the syntactic structure of a sentence can be:

- relative clauses,
- coordination markers,
- subordination markers,

- punctuation signs

Too much information to remember

This thesis introduces separately the "**too much information to remember**" issue as it can be expressed in a number of ways, some of them covering some of the already discussed TC issues, some not, but all posing a serious burden to human readers. It can be defined as having a large number of a particular type of sentence elements, which makes them difficult to remember, such as:

- large number of words in the sentence (nouns, verbs, adjectives, adverbs),
- large number of syntactic relationships,
- large number of modifiers of the same noun,
- different synonyms used for the same concept in the same sentence/text

The problem for human readers is, as mentioned, the need to keep all of the information in short-term memory while processing the rest of the sentence and of the text (Harley, 2008; Jurafsky and Martin, 2008) in an attempt to build their meaning. This TC issue can be measured by calculating the amount of any of the above described elements. While it is easy to calculate the number of words per sentence, calculating the other three measures is more difficult. In particular, calculating the number of synonyms used for the same concept in the text requires collecting all similar local contexts and comparing them, while calculating the number of modifiers per noun or syntactic relationships per sentence requires some pre-processing using a parser. The amount of syntactic relationships is related to the degree of syntactic complexity of the sentence, explained in the previous TC item.

Passive voice

The **grammatical voice** is a verb category which indicates whether the subject acts or is acted upon (Quirk, 1985). There are two forms of the grammatical voice – active voice, which indicates that the subject acts, and **passive voice**, which indicates that the subject is acted upon. An example follows below:

- Sentence A: *The girl wrote the letter.* (active sentence)
- Sentence B: *The letter was written by the girl.* (passive sentence)

The passive voice is considered to be cognitively more difficult to process than the active voice, because it is maintained that the reader needs first to transform the passive sentence into an active in order to understand it (Harley, 2008).

It is difficult to quantify the presence of passive voice in the text, because only the explicit constructions, marked by "*-ed by*", can be identified, which underestimates the actual number of passives (Quirk, 1985; Cohen et al., 2010).

Negative constructions

Negation can be defined as a contradiction or a denial of a word, clause, or sentence. In psycholinguistic research it is known that the comprehension of negative constructions takes more time to be understood by the reader, because he/she must first access the meaning of the affirmative version of the statement and then the negated one. (Glenberg, 1999). An example of this process (taken from Glenberg, 1999) is:

In order to understand the statement "*The buttons are not black.*" the reader must first process the sentence "*The buttons are black.*" and then the negation "*This is not true.*".

Computer applications also have problems with processing negation due to the difficulty in identifying it and determining its scope (Szarvas, 2008). The fact that they are problematic to identify constitutes a problem in quantifying their presence in text.

2.1.3.4. Discourse high text complexity issues

As explained in Table 2.1, the high discourse TC issues are illogical order and missing dicsourse connectives. A discussion of each follows below.

Pronouns with unclear reference

This thesis defines **pronouns** as words which can serve as substitutes for nouns, noun phrases, clauses and discourse segments (Quirk, 1985) and represent a closed class of words with nominal function. There are many different types of pronouns, such as: personal (*you, we, me, her, him*), reflexive (*ourselves, herself, myself*), possessive (*mine, his, theirs*), relative (*that, who, which*), interrogative (*which, whom, what*), demonstrative (*this, that, those*), indefinite (*each, any, somebody*), negative (*none, nobody, nothing*). As their function is to replace a word or an expression which has usually already been mentioned, the difficulty in processing them derives from the difficulty of identifying the noun phrase which they are referring to, i.e. their antecedent (Mitkov, 2002).

Both some human readers and computer applications can have difficulties processing pronouns and identifying the object to which they point. Particular categories of readers, for example patients

suffering from aphasia, cannot process pronouns (Canning, 2002). In NLP there exists a whole research area whose aim is to resolve anaphoric links between pronouns and other parts of speech (Mitkov, 2002). An example of a highly ambiguous pronoun is "*it*" in the following sentence:

"*Remove the bolt from the cover and slide **it** to the left.*" A shallow way to measure the amount of pronouns in text is to calculate them using a pre-defined list.

Illogical order

This thesis defines as **illogical order** an order of statements in a text which does not follow the usual cause-consequences principle. The illogical ordering of statements can result in risky or even fatal consequences if the texts are instructions which need to be executed (Heurley, 2001). An example taken from a medical protocol is:

"*Connect an empty 10ml luer-lock syringe and draw 5ml of blood (1ml if it's a new-born).*" (taken from MESSAGE Project).

These phenomena cannot be normally identified and addressed by computers, because they require extensive domain and world knowledge. For this reason, it is not considered possible to quantify the presence of this issue in text.

Missing discourse connectives

This thesis defines **discourse connectives** (Schiffrin, 1987), as words which connect two statements in order to express their relative logical and sequential ordering. Examples of discourse connectives are: "*First...*", "*Second...*", "*...then...*", "*Next…*", and "*Finally,...*".

Missing discourse connectives can affect human comprehension, as readers may not understand the relationships between statements or the order in which the events are happening or the order in which actions need to be taken, if the connectives are left implicit. An example of ambiguity created by the absence of discourse connectives can be seen in the following:

Sentence A: "*Today, I was walking in the park and listening to music*." can be interpreted in two ways:

1. The speaker is enumerating the activities he/she was doing today: first walking in the park, and then listening to music.

2. The speaker is telling about the unique event of simultaneously walking in the park and listening to music.

Compare sentence A to sentence B:

"*Today, I was first walking in the park and then listening to music.*"

The reader clearly understands from sentence B that the speaker was first walking in the park and then listening to music. Ambiguity due to lack of discourse connectives can be risky if the reader is not familiar with the domain of the text and has to follow precise instructions.

Calculating the presence in text of the amount of missing discourse connectives could be difficult. One of the ways to accomplish it is to use one of the results-intensive NLP techniques to follow cohesion chains and calculate the broken ones (Burstein et al., 2010). An easy solution can be to

calculate the number of existing discourse connectives, keeping in mind that more of them are there in a text, the easier to understand the text is. As it will be seen in Section 2.2., the highest level NLP techniques for detecting non-coherent texts require the use of latent semantic analysis (McNamara et al., 2010).

Text complexity factors concluding remarks

The previous sections have presented the concept of text complexity and a detailed analysis of the factors which affect it in the context of the English language. Although it has been proven that there is a clear improvement in text comprehension and in the performance of computer applications which process text with less complex texts, there are interesting interactions between some of the high text complexity issues. In fact, sometimes addressing some TC issues can introduce new ones making the text still difficult to read. An example is the relationship between word frequency and word ambiguity. In fact, although it is considered that more frequent words are better understood, according to the Zipf's law, the more frequent the words are, the more ambiguous they are as they are shorter and shorter words are more ambiguous (Zipf, 1949). Another interesting example is the relationship between the lexical word ambiguity and the length of sentences. It is known that from the syntactic point of view, longer sentences usually increase the complexity of the syntactic structure and the amount of information to remember, but on the other hand they also provide much more context necessary to disambiguate ambiguous words, while short sentences (which are otherwise considered to be simpler to understand) do not.

It is also important to specify that different text complexity issues affect the comprehension of different groups of readers differently, depending on their age, nationality, linguistic disorder, level of literacy, etc., and for this reason it is very difficult to provide an unique definition of a simple text or to design an approach to measuring text complexity or to text simplification which could be easily transferred to other kinds of readers. The following section will introduce the variety of related work in measuring TC.

2.2. Measuring Text Complexity

Measuring text complexity is important, as this process may help to determine which texts have to undergo simplification. The approaches to measure text complexity can be classified into two groups: early approaches and modern approaches. The early approaches correspond to the so called "**readability formulae**" and consist of counting in text the occurrences of a small subset of the earlier discussed text complexity markers (usually only surface markers) and produce a numerical value which can correspond either to the relative complexity level of a text, to a school grade level, or to the chronological age of the reader (Gunning, 1952; Flesch, 1948, Kincaid *et al.*, 1975; McLaughlin, 1969; Dale and Chall, 1948; Spache, 1953; Lorge, 1948; Yngve, 1960). The first readability formulae were originally created in the 1920s by educators in the United States (DuBay, 2004), with the aim of determining the complexity level of a text and to select texts appropriate for a particular school level. Hundreds of readability formulae have been developed since then. Although the main research on readability is for English, there also exist adapted readability formulae for many other languages, such as French, Spanish, German, Dutch, Russian, Swedish, Hebrew, Hindi, Chinese, Vietnamese and Korean (Rabin, 1988).

The existing modern approaches usually analyse a significantly higher number of TC measures and produce series of several numerical measures of its complexity (McNamara et al., 2010). The approaches which are still based on a small set of surface high text complexity markers either use them to produce a measure of a restricted type of text complexity (e.g. syntactic complexity, Szmrecsanyi, 2004), or apply machine learning techniques to automatically predict the general linguistic complexity of a text (Van Oosten et al., 2010; Nenkova et al., 2010). Some approaches, also study the correlation of the existing readability formulae with human ratings of text difficulty (Van Oosten et al., 2010; Leroy, 2010).

The summary of all the main existing approaches is provided in Table 2.2. The approaches are listed from the oldest to the most modern ones and details about each of them are provided.

Approaches	Reference	Unique score	Counting TC markers	Machine learning	Human ratings
Early approaches	Flesch, 1948	school grade level, reading age	ASL, AWL-S	N/A	N/A
	Dale and Chall, 1948	school grade level, reading age	ASL, PDW	N/A	N/A
	Lorge, 1948	reading age	ASL, NPPh, PDW	N/A	N/A
	Gunning, 1952	school grade level, reading age	ASL, P3SW	N/A	N/A
	Spache, 1953	school grade level, reading age	ASL, PDW	N/A	N/A
	Yngve, 1960	sentence complexity	PTD		
	McLaughlin, 1969	school grade level, reading age	P3SW	N/A	N/A
	Kincaid et al., 1975	school grade level, reading age	ASL, AWL-S	N/A	N/A
Recent approaches	Szmrecsanyi, 2004	sentence syntactic complexity	NSC, NWh, VP, NP	N/A	N/A
	McNamara et al., 2010	N/A	200 measures	N/A	N/A
	Van Oosten et al., 2010	school grade	ASL, AWL-S, PDW, PSW, TTR, P3SW, PW6Ch	yes	yes
	Nenkova et al., 2010	N/A	ASL, PTD, APhL, PP, NP, VP, PhL, NHNM	yes	N/A
	Leroy, 2010	yes	several readability formulae	N/A	yes

Table 2.2: Overview of the approaches for measuring text complexity.

As Table 2.2. shows, most of the early approaches rely on counting high text complexity markers,

while the recent approaches on machine learning. In the table, ASL stands for average sentence length, calculated in number of words, AWL-S stands for average word length calculated in syllables, P3SW is the percentage of words with 3+ syllables, PDW is the percentage of difficult words from a manually composed list, NPPh is the number of prepositional phrases per 100 words, PTD stands for parse tree depth, NSC – number of subordinate conjunctions, NWh – number of "*wh-*" pronouns, VP – number of verb phrases, NP – number of noun phrases, PP – number of prepositional phrases, PSW – percentage of sentences per word, TTR – type/token ratio, PW6Ch – percentage of words with 6+ characters. APhL – average phrase length, PhL – average phrase length, and NHNM – number of head noun modifiers.

The approach presented in this thesis is similar to the approaches counting the number of high text complexity issues (McNamara et al., 2010 and the classic readability formulae), but differs from the classic readability formulae in not generating an unique score. Finally, it differs from all the approaches as they are not crisis management domain-specific. The approach chosen in this thesis for evaluating TC of Crisis Management texts will be presented in Chapter 3.

2.3. Reducing Text Complexity

This thesis proposes text simplification as a method to reduce text complexity. This section will present the existing approaches. Specifically, Section 2.3.1 will define the concept of text simplification and provide a short overview of the approaches. Section 2.3.2 will present the manual text simplification approaches, Section 2.3.3 will present the semi-automatic approaches and Section 2.3.4. – the fully automatic text simplification approaches.

2.3.1. Text simplification definitions and overview of the approaches

Text simplification can be defined as any process which reduces the syntactic or lexical complexity of a text while attempting to preserve its meaning and information content (Siddharthan, 2003). Gasperin et al. (2009b) define text simplification as a "research area of Natural Language Processing, whose goal is to maximize text comprehension through simplification of its linguistic structure". This thesis defines text simplification as a manual/semi-automatic or automatic approach to reducing text complexity at different levels (lexical, syntactic, or discourse), while maintaining the information contained in it. More about what constitutes text simplification can be understood from the following overview of existing text simplification approaches.

There exist various approaches to text simplification, including manual approaches, called 'controlled languages', and semi-automatic or fully-automatic text simplification tools. A distinction will be made according to whether they employ the controlled languages approach or not, and whether they are orientated towards increasing text comprehension for human readers or towards facilitating text processing of computer applications. Among the fully automatic systems, a distinction will be made according to whether the systems are independent (i.e. their one and only purpose is to simplify text), or integrated in other systems.

2.3.2. Manual simplification: Controlled languages

"Industry does not need Shakespeare or Chaucer, industry needs clear, concise communicative writing - in one word, Controlled Language." (Goyvaerts, 1996)

To the best of our knowledge, the only manual text simplification methods are those, based on "Controlled Natural Languages" (CNL) or "Controlled Languages" (CL). CLs are guidelines containing sets of rules for manual simplification of complex texts or for writing simple texts from scratch. The term "controlled language" was first used to refer to the restrictive set of rules set by Charles Ogden in 1930 in his book "Basic English: A General Introduction with Rules and Grammar". There exist several CL definitions. Most of the definitions of controlled languages focus on two different points of view of the CNLs: either on the fact that CLs restrict language at different levels in order to reduce its ambiguity and complexity (Hoefler and Bunzli, 2009; Angelov and Ranta, 2009; Kuhn, 2009b; AECMA, 1995; Gough, and Way, 2003), or on the fact that controlled languages are a subset of natural language (Kuhn, 2009a; Angelov and Ranta, 2009; Wyner et al., 2009; Gough and Way, 2003) which has behaviour similar to natural language. The second definition creates some overlap between the term "sublanguages" and the term "controlled languages", which this thesis does not consider to be the case.

This thesis defines controlled language as a set of writing rules which poses artificial restrictions on the use of the natural language on different levels (lexical, syntactic, discourse) in order to create simple-to-comprehend texts or texts which are easy to process by computer applications. Controlled

language can be applied both on the general language and on sublanguages, with application to sublanguages allowing better precision since their language is already naturally restricted. In this respect, this thesis agrees with Kittredge (2003) and underlines the fact that the difference between controlled languages and sublanguages lies in the fact that controlled languages are artificially restricted, while sublanguages arise naturally in highly specialised communication. The aim of the controlled languages approach is to avoid or at least to reduce text complexity and ambiguity. Thus, it is a good method for avoiding producing texts which are difficult to understand, or which could be misunderstood by less competent language users. This is achieved by keeping under control all text complexity issues affecting the specific kind of text. Spiaggiari et al. (2005) defines the aims of controlled languages as being to simultaneously ensure readability (which he defines as simplification of syntactic structures), comprehensibility (considered to be substitution with more understandable lexical terms), and translatability (in order to ease the transfer between languages) of text. It is considered that the aim of different controlled languages should be different, according to their final purpose and the domain of application.

The need for controlled languages was first seen in teaching English to non-native speakers, as they have a limited vocabulary and cannot easily process sentences with complicated syntax. Basic English (Ogden, 1930) is an example of a controlled language which is used to address this aim. Lately, CLs have been developed mainly for the technical field for writing technical documentation. The most famous example of such CLs is the former AECMA (Association Européenne des Constructeurs de Matériel Aérospatial), now ASD-STE 100 (Aerospace and Defense Industries Association of Europe Simplified Technical English) (AECMA, 1995), which is widely used in the domain of aeronautics and was introduced in order to avoid fatal accidents due to misunderstanding, such as the Tenerife air crash (Air Line Pilots Association, 1977). Other examples include Xerox, Boeing, Rolls-Royce, Saab, General Motors, and IBM.

There are CNLs which are designed to simplify and make uniform many different types of texts. Controlled languages are used in the scientific field for mathematical proofs (Cramer et al., 2009), in the legal field for contracts and legal forms (Pace and Rosner, 2009), in the medical field for writing clinical practice guidelines (Shiffman et al., 2009), etc. Concrete examples include: Controlled Legal German (Hoefler and Bunzli, 2010), used to facilitate semantic processing of Swiss statutes and regulations; CNLs for the Semantic Web, which can be translated into Web Ontology Language – Rabbit (Hart et al., 2008), CLOnE (Funk et al., 2007), Lite Natural Language (Bernardi et al., 2007), and Controlled Language for ANNotation (Dantuluri et al., 2010), which is used as an interface to Semantic Web applications. CLANN allows novice users to edit and annotate project documents, thus making these documents easy to be parsed for extracting implicit knowledge; Naproche CNL, a controlled language for authoring of mathematical proofs, whose main aim is to "make formal mathematics more readable to the average mathematician" (Cramer et al., 2009); economic and business CNLs, like the Economical Discourse Representation Theory, based on Combinatory Categorial Grammar (Bos, 2010) and the Semantics of Business Vocabulary and Business Rules which is based on formal logic (Spreeuwenberg, and Andreson Healy, 2009); Discourse-Based Reasoning (Potter, 2009), based on natural discourse and argumentation theory and Rhetorical Structure Theory (Mann and Thompson, 1988). Although CNLs are developed mainly for English, there exist examples also for other languages such as Esperanto, French (Barthe, 1996, 1998; Cardey, 2011; Renahy et al., 2010), German (Schactl, 1996), Swedish (Alqvist et al., 1996), Spanish (Bustamante, 2000), Japanese, Chinese (Zhang, 1998), Mandarin (Pool, 2006), Modern Greek (Vassiliou et al., 2003), Spanish (Blanco 2009) and Polish (Bogacki, 2009; Rudas 2009), with prototypes existing also for Bulgarian (Temnikova and Margova, 2009).

According to (Arnold et al., 1993), CNLs can be human-orientated or machine-orientated. Human-orientated CNLs aim to improve human communication (Wyner et al., 2009), (e.g. improve

comprehensibility of written texts for human readers, improve translatability of written texts). In contrast, machine-orientated CNLs aim to improve human-machine communication (Wyner et al., 2009), and in this way, automatic processing, and thus either to allow easy mapping of the controlled text to a formal logic (like the Attempto Controlled English, ACE, which can be translated automatically and unambiguously into logic (Kuhn, 2009b), or to improve the performance of computer applications, such as for example machine translation engines. Another important difference between human-orientated and machine-orientated CNLs is that human-orientated CNLs have freely defined general rules, such as “Use short sentences”, “Avoid passive voice” and “Avoid pronouns”. In contrast, machine-orientated CNLs, also called logic-based controlled languages, are well defined and thus allow mapping to an existing formal language, are easily transferable to other knowledge representation languages, and allow automatic consistency checks. Examples of human-orientated CNLs are Basic English (Ogden, 1930) and Plain English, while examples of machine-orientated controlled languages include PENG (Pool, 2006); Xerox's controlled language, which is both aimed at improving human comprehension and machine translation of technical documents (Ruffino, 1982); and ACE, used for a variety of applications, including ontology searching tools and automatic text summarisation (Kuhn, 2007). These CLs will be described in more detail further on.

The subdivision can also be defined also to be in formalism-like systems with strict syntax and semantics and rich fragments of natural language, informally specified by heuristic rules (Angelov and Ranta, 2009). The formalism-like systems are usually those which address computer applications, as they are easier to implement, while the heuristic-rules-based ones are usually addressing human readers.

Controlled languages can be also classified according to the types of rules they feature. The rules can be either prescriptive (listing the allowed expressions), or proscriptive (specifying the disallowed structures and terms). Another classification of the kinds of rules used in controlled languages in order to reduce natural language ambiguity and complexity is provided in Hoefler and Bunzli (2009). According to their classification, there are the following types of rules:

1. Construction rules
2. Interpretation rules
3. Paraphrases

Different CLs feature some or all of these kinds of rules. Construction rules restrict the lexical and syntactic expressions allowed to be used in controlled language texts. Interpretation rules assign fixed interpretations to the allowed ambiguous lexical and syntactical expressions. The third type of rules, the paraphrases, suggest alternatives to the forbidden lexical and syntactic expressions.

Although controlled languages are a manual text simplification approach, tools to support writing texts according to them or to check consistency of already-written texts have been developed lately. An overview of the semi-automatic and fully-automatic approaches to text simplification will follow in the next sections. Table 2.3 shows the distribution of controlled languages among the manual, semi-automatic and fully-automatic text simplification approaches.

Simplification approaches	**CL-based**	**Not CL-based**
Manual	YES	Not reported
Semi-automatic	YES	YES
Fully Automatic, integrated in other NLP applications	YES	YES

Independent Fully Automatic	NO	YES

Table 2.3: Controlled languages and text simplification approaches

The first column of the table lists the different types of approaches, starting from the manual and ending with the fully automatic ones, while the second column gives an indication of whether there exist any of these types of approaches which are controlled language-orientated. The third column indicates whether there are any TS approaches which are not controlled language-orientated. “YES” indicates that there exist such approaches, “NO” that there are no such approaches, and “not reported” that there is no information about that.

Some examples of human-orientated, machine-orientated and mixed-purpose controlled languages will be presented in Sections 2.3.2.1, 2.3.2.2 and 2.3.2.3, respectively.

2.3.2.1. Human-orientated CLs

Human-orientated CNLs address the natural language comprehensibility for different kinds of audiences – non-native speakers (Basic English, Ogden, 1930), non-competent readers (Plain English Campaign), communication in emergency situations (PoliceSpeak, Johnson, 1993 and SeaSpeak, Strevens, 1984). These examples of human-orientated controlled languages are described in detail below.

Basic English

Basic English was created by Charles Kay Ogden in 1930 and consists of a vocabulary composed of the most frequent and familiar 850 English words. Basic English is considered an example of "international auxiliary languages" (IAL), which are used for communication between people from different nations. In this role, Basic English became very famous in the context of the Second World War, as it was promoted as a tool for world peace (Ogden,1930). The list of 850 words can be further expanded to 1500 words with 350 international words and 300 words for the fields of science and economics, or to 2000 words, considered enough for a "standard English level". The list of 850 words can be found at http://en.wiktionary.org/wiki/Appendix:Basic_English_word_list[7]. The limitation of this CL is that it reflects the language of the time, and new studies should be conducted in order to determine the currently most frequent words; also, no objective evaluation of its performance in facilitating comprehension has been reported.

Plain English Campaign

Another example of a controlled language aimed at simplifying texts for human readers is Plain English, which is promoted by the Plain English Campaign. Plain English is a set of guidelines for use by public services (government, local councils, banks, medical staff, and insurance companies) with the aim of teaching them how to write their documents in an accessible language which is orientated towards the target audience. The rules are very general and can be found in the guide "How to write in Plain English". Some examples are presented below:

- Keep your sentences short,

7 Last accessed on November 14th, 2011.

- Prefer active verbs,
- Use “you” and “we”,
- Choose words appropriate to the reader

The Plain English campaign has also produced glossaries containing accessible alternatives to general or domain-specific (law/medical/financial) terms. Some examples from the lists of general and medical terms are given in the tables below. The terms in bold are the suggested alternatives.

General terms:

Accelerate	**Speed up**
Accentuate	**Stress**
Accommodation	**Where you live, home**
Accompanying	**With**

Medical terms:

Amnesia	**Loss of memory**
Analgesic	**Something that lessens pain**
Anastomosing	**Joining together**
Aneurism	**A swelling in an artery**

Examples of sentences, re-written according to Plain English rules and alternative terms, can be found on the Plain English Campaign website. An interesting example is the sentence provided below:

- Original sentence: "*If there are any points on which you require explanation or further particulars we shall be glad to furnish such additional details as may be required by telephone.*"

- Re-written sentence: "*If you have any questions, please phone.*"

The limitation of the Plain English Campaign is that since it addresses manual re-writing, it features very generally defined rules which cannot be easily implemented, and also a very small number of rules. Furthermore, no objective evaluation has been reported.

PoliceSpeak and SeaSpeak

PoliceSpeak (Johnson, 1993) has been funded by British Telecom, the Home Office, and the Kent County Council and developed in the context of the Channel Tunnel to facilitate and make unambiguous the communication between British and French officials. It was developed on the basis of a corpus analysis of police communications and thus it takes into account the lexical and syntactic particularities of the police language, the domain situations of managing the Channel Tunnel, and the particularities of translation between English and French of such a restricted sublanguage.

SeaSpeak (Strevens 1994) was developed by the International Maritime Lecturers Association (IMLA)[8] after the disaster involving the ferry Scandinavian Star (Solheim et al, 1992), which was due to lack of common language between the crew members and killed over one hundred fifty people. SeaSpeak in its current form, Standard Marine Communication Phrases (SMCP)[9], is the specialised language used by the international community at sea. It is composed of a set of fixed short phrases, sometimes also featuring non-English words.

The limitations of these two controlled languages are that they are concerned with very restricted sublanguages and situations. Both PoliceSpeak and SeaSpeak have been developed following the development of the CL for air industry communication (AECMA or ASD-STE 100), which is described in Section 2.3.2.3.

8 International Maritime Lecturers Association (IMLA). Last accessed on November, 14th, 2011.
9 http://www.imo.org/OurWork/Safety/Navigation/Pages/StandardMarineCommunicationPhrases.aspx. Last accessed on November 14th, 2011.

2.3.2.2. Machine-orientated CLs

Machine-orientated CNLs allow easy mapping of text to formal languages or formal knowledge representations and thus are often connected to reasoning engines (Angelov and Ranta, 2009). A formal language is "a set of strings, each string composed of symbols from a finite symbol-set called an alphabet" (Jurafsky and Martin, 2008) and which can be described using formal grammars, such as for example finite automata (Mateescu and Salomaa, 1997). The most famous examples of formal controlled languages are PENG (Schwitter, 2008) and ACE (Kuhn, 2007), while minor ones are the CLs for mathematical proofs (Cramer et al., 2009); the CLs for the Semantic Web (Rabbit, Hart et al., 2008), CLOnE (Funk et al., 2007), Lite Natural Language (Bernardi et al., 2007), and CLANN (Controlled Language for ANNotation, Dantuluri et al., 2010); and economic and business CLs (Bos, 2010; Spreeuwenberg and Andreson Healy, 2009). PENG and ACE are described below.

Processable ENGlish (PENG)

Processable English (Schwitter, 2008) is used for writing technical specifications and possesses a restricted lexicon and a restricted grammar. The lexicon features a set of pre-defined function words, a list of disallowed words, and a list of predefined allowed content words. The user can expand the lexicon by adding new words, as well as by defining synonyms or acronyms. The controlled language grammar defines the allowed structure of simple sentences and the ways that simple sentences can be combined into complex ones. A definition of simple PENG sentences follows below:

Sentence -> Subject + Predicate
Subject -> Determiner

```
   {+ Pre-nominal Modifier}
   + Nominal Head
   {+ Post-nominal Modifier}
Subject -> Nominal Head
Predicate -> { Negation }
   + Verbal Head
   + Complement
   {+ Adjunct}
```

Examples of allowed simple sentences matching this pattern are provided on the PENG website[10], and are:

- The butler works.
- The butler works in Dreadsbury Mansion.
- The mother of the butler does not work in Dreadsbury Mansion.
- Every butler hates a person.
- No person hates every person.
- Agatha hates Charles or the butler.
- Agatha is not identical to the butler.
- Butlers are murderers.

10 http://web.science.mq.edu.au/~rolfs/peng/, last accessed on November 14th, 2011.

Attempto Controlled English (ACE)

Attempto Controlled English (Kuhn, 2007) has been used for software specifications, theorem proving, automatic text summarisation, ontologies, and other applications. Like PENG, Attempto Controlled English has a predefined vocabulary featuring function and content words. The vocabulary is expandable. The ACE's grammar defines the meaning of ACE texts. For example, “every” is interpreted as “universally quantified”, like in the sentence “Every cat has a tail.”. ACE allows both simple and complex sentences, with complex sentences being composed of simple ones. The allowed combinations of simple sentences are through coordination, subordination, quantification, and negation. Only two forms of questions (called “queries”) are supported: yes/no queries and wh-queries. Ambiguity is handled by either avoiding ambiguous statements, or by providing pre-defined, unique interpretations for some ambiguous statements. For example, the sentence:

“*A customer inserts a card which is valid and opens an account.*” which in natural language is ambiguous and can mean both

1) “*A costumer inserts a valid card. The card the costumer inserts opens an account.*” and

2) “*A costumer inserts a valid card and then the same costumer opens an account.*”

According to ACE, this sentence would have only one meaning, corresponding to (2). Meaning one is achievable only if “that” is inserted. Although anaphoric references are generally considered a text complexity issue, in ACE anaphoric references expressed through pronouns are allowed.

The limitations of these CLs are that they are very domain and document specific and cannot be transferred to any other domains. The tools designed to work with these two CLs will be presented in Section 2.3.3.1.

2.3.2.3. Mixed-purpose CLs

As mentioned before, one of the controlled language definitions implies that controlled languages ensure readability, comprehensibility and translatability of the text (Spiaggiari et al., 2005). This means that some of the controlled languages aim to ensure both comprehensibility for human readers and ease of text processing for computer applications (the most frequent case being Machine Translation engines). Examples of such controlled languages are some of the CLs for technical documentation, which aim to facilitate the translation of their documents in addition to making their documents easy to read and find information in. In fact there exist, as it will be seen later, MT tools designed specifically for a particular controlled language (see Section 2.3.4.), but details about their CLs are not published due to confidentiality. Publicly available information exists about the ASD-STE 100 (formally AECMA) and LiSe controlled languages.

ASD-STE 100 (Simplified Technical English)

Created by the European Association of Aerospace Industries (formally AECMA), ASD-STE 100 is a specification for writing aircraft documentation. The specification is known formally as AECMA Simplified English (SE) (AECMA, 1995). ASD-STE 100 is designed for manually writing documentation and has a set of writing rules and a dictionary which poses limitations on the number of words, the meanings, and the possible POS with which they have to be used. The rules are similar to the Plain English Campaign, e.g. "Use short sentences.", "Avoid passive voice.", but there are also more document-specific rules, e.g. "Introduce a list item with a hyphen" (Unwalla, 2004). In this respect ASD-STE 100 is similar to the CL presented in Chapter 4, but it is specific for the aircraft domain and documents.

The Controlled Language "LiSe"

LiSe has been designed both for improving human comprehension of crisis management texts and facilitating machine translation for the French language. LiSe is created on the basis of a long collaboration of the Centre Tesnière[11] with specialists from the aerospace domain, hospitals, and Crisis Management government centres, and on the basis of a preliminary manual analysis of a corpus of collected documents, which are mainly protocols from the healthcare domain (Renahy et al., 2011). The aim of LiSe is to help healthcare specialists to write clearly understandable documents by importing into their field the CLs which existed for a long time in the field of technical documentation. During MESSAGE project[12] the controlled language philosophy of LiSe has been transferred to three other European languages: Spanish (Blanco, 2009), English (Temnikova and Orasan, 2009) and Polish (Cholewa, 2009; Gwiazdecka, 2009; Rudas , 2009).

11 http://tesniere.univ-fcomte.fr/. Last accessed on November 14th, 2011.

12 Full title: Alert Messages and Protocols, project financed by the European Union (JLS/2007/CIPS/022). With the support of the Prevention, Preparedness and Consequence Management of Terrorism and other Security-related Risks Programme European Commission - Directorate-General Justice, Freedom and Security.

Some research has also been conducted towards the development of CL prototypes for the CM domain for Modern Greek (Papadopoulou and Puig Portella, 2009) and Bulgarian (Temnikova and Margova, 2009).

LiSe has been developed for protocols for healthcare specialists (Renahy et al., 2011), but has also been extended to quick reaction documents, such as alerts and short messages. It has guidelines containing rules for the structuring of information and formatting rules, but also concrete rules at the syntactic and lexical levels. There are rules adapted for the particular document structure, the application domain, and the target audience, as well as the target languages in the case of eventual manual or automatic translation (Renahy et al., 2011). The source language from which the translation is made is therefore a French Controlled Language, i.e. the Controlled Pivot Source Language (CPSL, Cardey, 2011). The output languages for translation from French targeted in the LiSe Project are English, German, Chinese, Thai, and Arabic. A writing aid for texts in CLs has also been developed in the context of the project LiSe (Renahy et al., 2010, discussed in Section 3.3.1.). The research team is also working towards the development of a rule-based MT system tailored for their specific CL and domain, in order to avoid any fatal consequences related to wrong translations of documents for emergency situations (Cardey, 2011). The shortcoming of this CL is that it exists only for French, so it needs to be transferred to English in order to be adapted for the purposes of this thesis.

2.3.3. Semi-automatic or computer-aided text simplification

Mitkov (2007) distinguishes Computer-Aided Language Processing (CALP) from general NLP and underlines that in it, language processing is not done entirely by computers, but rather human intervention improves, post-edits or validates the output of the computer program. Several CALP

systems exist in different NLP areas, including:

- Machine-Aided Translation (Kay, 1980)
- Computer-Aided Text Summarisation (Orasan, Mitkov, and Hasler, 2003)
- Computer-Aided Generation of Multiple-choice texts (Mitkov, Ha, and Karamanis, 2006)
- Computer-Aided Information Extraction (Cunningham, 2002)
- Annotation of corpora (Orasan, 2005)

Similarly, this thesis defines Computer-Aided text simplification as text simplification which is assisted by the computer, but in which the process is not totally automatic, i.e. it requires human intervention. This thesis divides Computer-Aided text simplification systems into those which are controlled language-based, and those which do not involve a controlled language. The controlled language-based tools can be controlled language editors (Renahy et al., 2010; Schwitter et al., 2003), and controlled language grammar checkers (Mitamura and Nyberg, 2002). The controlled language editors and spell-checkers aim at simplifying the work of the controlled language text writer, because it is known that writing text according to controlled language rules can be a cognitive-effort-intensive and very time-consuming process (Goyvaerts, 1996; Huijsen, 1998). The controlled language editors either offer choice of structures with possible expressions and lexical terms (Renahy et al., 2010) or are predictive and automatically continue the sentence while typing (Schwitter et al., 2003). The first steps towards implementing a controlled language writing aid, will be presented in Chapter 7.

The non-CL-based tools are also grammar checkers and can be classified into two categories: tools which only identify problematic issues in text and signal them to the user (Graesser et al., 2006), and tools which identify the problematic issues and additionally suggest alternatives for the user to choose from during the post-editing process (Liben-Nowell, 2000; Max, 2005). A third type of

semi-automatic text simplification tools, which can be both controlled language-orientated and not controlled language-orientated involve the automatic recording of simplification operations for subsequent use, such as for the purposes of machine translation and building of parallel corpora (Renahy et al., 2010; Caseli et al., 2009). The limitations of these systems are that either they are very domain-specific (in case of the controlled language-based tools), or they address only a restricted number of TC issues, probably due to implementation difficulties, while the rest are left to be manually re-written by the user.

Reference	Editor	Offers choice of structures	Typing automatic prediction	Grammar checker	Highlights problematic issues	Highlights problematic issues & offer a re-writing alternative	Application
Controlled language-based tools							
Mitamura and Nyberg, 2002	N/A	N/A	N/A	yes	N/A	yes	Pre-processing the machine translation system input
Schwitter et al., 2003	yes	N/A	yes	N/A	N/A	N/A	PENG controlled language, technical specifications
Renahy et al., 2010	yes	yes	N/A	N/A	N/A	N/A	LiSe controlled language - crisis management and medical leaflets
Not controlled language-based tools							
Liben-Nowell, 2000	N/A	N/A	N/A	yes	N/A	yes	Assisting in the writing of texts suitable for readers suffering from aphasia (Powell, 2010)
Max, 2005	N/A	N/A	N/A	yes	N/A	yes	Assisting in the writing of texts suitable for readers suffering from aphasia (Powell, 2010)
Graesser et al., 2006	N/A	N/A	N/A	yes	yes	N/A	Assisting writing clear survey questions
Caseli et al., 2009	yes	N/A	N/A	N/A	N/A	N/A	Building a Brasilian Portuguese parallel corpus of complex and simplified texts

							for learning operations for automatic text simplification

Table 2.4: Overview of semi-automatic TS approaches

Table 2.4 lists first the controlled language-based semi-automatic approaches and then the not controlled language-based ones in chronological order. Due to the fact that the existing computer-aided text simplification tools are restricted to their specific application domains (column 6 in Table 2.4), Chapter 7 will propose the first steps towards the implementation of a semi-automatic writing tool in accordance with the controlled language described in Chapter 4. In addition, as nothing has been reported about studying the user requirements before implementing the CL-editors described in Table 2.4, Chapter 7 will also present the first documented investigation of the priorities in implementing a controlled language writing aid.

2.3.4. Fully-automatic text simplification systems

This thesis defines fully-Automatic Text Simplification systems (f-ATS) as systems which do not require human intervention and translate complex text into simple text on their own. There are two different kinds of fully automatic systems – those performing text simplification as their one and only purpose, and those which integrate text simplification as a pre- or post-processing step of another NLP application with different end purposes. To the best of our knowledge, there is no existing independent f-ATS system which is based on the controlled languages approach, while among the non-independent f-ATS systems there are some (e.g. the ones integrated in a Machine Translation system). This is another difference between the independent and non-independent f-ATS. Below follows a description of the independent f-ATS systems, and then a description of automatic text simplification integrated into other NLP tasks.

2.3.4.1. Independent f-ATS

Relatively few independent f-ATS systems have been developed, but none for controlled languages. Independent f-ATS systems can be classified according to two inter-related criteria: purpose, and coverage. According to the purpose of simplification, fully automatic systems can be divided into those which address text complexity reduction for human readers (Devlin, 1999, Canning, 2002, Siddharthan, 2003, Inui et al., 2003, Gasperin et al., 2009) and those addressing text complexity reduction for computer applications (Chandrasekar, 1996, Siddharthan, 2003). Both of the last two systems simplify text in order to facilitate a parser's work, while the systems addressing human readers focus on text simplification for:

- readers suffering from aphasia - (Devlin, 1999, Canning, 2002)
- deaf readers - (Inui et al., 2003)
- illiterate readers - (Gasperin et al., 2009)

From the point of view of coverage, f-ATS systems can be classified according to the level at which simplification occurs: lexical (Devlin, 1999, Gasperin et al., 2009), syntactic (Chandrasekar, 1996, Canning, 2002, Siddharthan, 2003, Gasperin et al., 2009) or discourse (Siddharthan, 2003). At the moment there is no system addressing all three levels of simplification, and the lists of text complexity issues addressed by these systems at any level are not exhaustive. In relation to the level coverage, in the best case, two levels have been combined: Siddharthan (2003) has addressed syntactic simplifications and their implications for discourse, while Inui et al. (2003) carried out both syntactic and lexical simplifications for the specific needs of deaf readers; Gasperin et al., 2009 also carried out both lexical and syntactic simplification, this time for readers with low

literacy levels. Although Gasperin et al. (2009) carried out lexical substitution of discourse markers, this is considered as being lexical simplification and not discourse simplification, as the rhetorical relations between predicates and text structure remained unchanged.

Text simplification at the discourse level has been carried out only by one of these studies and only in the last stage of text simplification as something like a post-processing stage. Siddharthan (2003) applies this kind of simplification because he maintains that performing syntactic simplification by splitting long sentences into shorter ones may cause alterations in the chronological order of events, may break anaphoric links, and may produce incoherent text if the necessary connectors are not generated. An overview of the TC issues addressed by each system is given in Table 2.5.

year	authors	NLP	human readers	language	lexical	syntactic	discourse
1996	Chandra sekar and Srinivas	parser (pre-processing step)	N/A	English	N/A	-subord. cl., -coord. cl., -relative cl., -appositions	N/A
1999	Devlin	N/A	aphasics	English	less frequent terms <- more frequent synonyms	N/A	N/A
2002	Canning	N/A	aphasics	English	N/A	-passive, -compound sent., -anaphora	N/A
2003	Siddhart han	parser	general: aphasics, deaf readers, non-native speakers, etc.	English	N/A	-subord. cl., -coord.cl., -relat. cl., -apposit.	-sentence order, -selection of cue-words, -refer. expr. generat.
2003	Inui et al.	N/A	deaf readers	Japanese	reducing lexical variety to 2000 words, more personal pronouns	learning paraphrases from aligned sentences, annotated by teachers of deaf students	N/A
2009	Gasperi n et al.	N/A	readers with low literacy levels	Brazilian Portuguese	Replace lexical terms with	Split sentences, passive → active,	N/A

					more common ones	reordering to SVO order, clauses reordering	

Table 2.5: Schematic history of independent f-ATS systems.

The first column of the table provides the year in which the system was created (**year**), the second column provides the authors of the system (**authors**), the third column provides the targeted NLP application (**NLP**), the fourth column provides the targeted human readers (human **readers**), and the remaining columns provide the linguistic coverage (**lexical, syntactic** and **discourse).**

2.3.4.2. Not-independent f-ATS

While only a restricted number of independent f-ATS systems exist, text complexity reduction from the specific point of view of particular applications has been applied in other areas of Natural Language Processing, resulting in many TS systems integrated into other applications. These areas include:

- **Text Reduction for Small Screens**: intended to fit text for use on small screens, e.g. mobile phones or subtitles (Daelemans et al., 2004, Corston-Oliver, 2001, Euler, 2002 and Grefenstette, 1998)

- **Text Summarisation**: most of the time text simplification is applied as a pre-processing step in order to remove redundant information (Siddharthan et al. 2004, Daume and Marcu, 2005a, Daume and Marcu, 2005b, Dunlavy et al., 2003, Vanderwende et al., 2007, Dorr et al., 2003, Knight and Marcu, 2002), but sometimes it can also be used as a post-processing step in order to adapt the summaries to particular readers (Elhadad, N., 2006; Elhadad and Robin, 1992; Lal and Ruger, 2002)

- **Information Extraction**: TS is used as a pre-processing step removing unimportant information, in order to transform the sentence to a sentence from which information can be more easily extracted. These applications do not apply TS to sentences which do not contain important information. (Klebanov et al., 2004, Jang et al., 2006)

- **Machine Translation**: text simplification and more particularly, Controlled Language-based simplification is used as a pre-processing step. Examples of such systems are the TITUS system (Streiff, 1985), which restricted the input syntax and vocabulary for machine translations in the textile industry; the XEROX system (Ruffino, 1982), which pre-edited their texts with their own controlled language rules; and the currently-under-development rule-based MT system by Cardey (2011) specifically for the Crisis Management domain.

- **Natural Language Generation**: applies text simplification methods to generate texts for low-skilled readers. An example is the SkillSum system (Williams and Reiter, 2009), which adapts literacy reports to readers with restricted reading skills.

- **Semantic Role Labelling**: uses TS as a pre-processing step. An example of such a system is Vickrey and Koller (2008), who split sentences into short ones in order to facilitate the task of assigning semantic roles to verb arguments.

- **Other applications**: an example being the system of Dras (1999), which transforms text in order to adapt it to conference requirements for length, readability, lexical density, and

sentential variation.

A typical factor of these applications is that there is very often information loss, because the type of simplification which they apply involves removing portions of texts, from single words to whole sentences. The simplification is usually done at all levels, depending on the system, but the list of TC issues is again not exhaustive, and the proposed simplification operations are usually specifically tailored to the targeted computer applications.

2.4. Conclusions

The purpose of this chapter was first of all to introduce the problem of Text Complexity (TC) treated in this thesis, and secondly to propose a solution to it. For this reason, a definition of text complexity together with an overview and a comparison of the different TC issues affecting human written comprehension and the performance of computer applications processing text has been provided. Both human readers' comprehension and computer application processing of text have been taken into consideration, because the aim of this thesis is to improve both through the proposed solution. It has been shown that although the reasons differ, many of the TC issues are common for both human readers and computer applications processing text, and thus need to be addressed, regardless of the which is the goal. The calculation of most of these features for evaluating the linguistic complexity of written text will be shown in Chapter 3. An overview of the existing approaches to measuring TC has been provided; this is necessary background to the TC

analysis of the domain texts presented in Chapter 3.

The second purpose of the Chapter was to introduce text simplification as the best solution to text complexity, which preserves original text's content. A definition of text simplification, together with an extensive overview of the existing approaches, ranging from manual to fully automatic solutions, has been provided. The TS approaches have been classified according to several different criteria: coverage of linguistic levels (lexical, syntactic, or discourse), purpose (human readers or computer applications), degree of automation (manual, semi-automatic, or fully automatic), controlled language-relatedness (yes or no). It has been shown that a range of readers (aphasics, deaf readers, non-native speakers, readers with low literacy levels) and a range of computer applications (small screen devices, text summarisation, information extraction, natural language generation, semantic role labelling and other applications) can benefit from text simplification.

As has been seen, the existing approaches have several limitations. This thesis would like to focus on those which make them particularly inappropriate for its purposes:

1. The semi- and fully-automatic TS systems are not exhaustively addressing the existing TC issues, even within their own restricted domain, but rather just a small subset of them, probably due to the difficulty of implementation.

2. Their implementations are also often not based on psycholinguistic findings.

3. They also very often involve information loss.

4. In most of the cases, proper testing with end-applications or readers has not been performed.

5. None of the existing approaches is tailored for the Crisis Management domain in particular.

All of these issues can be crucial in an emergency situation. For this reason, in order to apply text simplification or generation of simple texts approach in the Crisis Management domain, this thesis chooses a controlled language-based approach, because controlled languages can more easily be tailored to address more exhaustively the TC issues which are critical for the Crisis Management domain. As was seen in Section 2.3.2.3, the best approach to be adapted for the purposes of the thesis is the controlled language LiSe (Renahy et al., 2011), which needs to be adapted from French to the English language.

The next chapter will present the Crisis Management domain and the text complexity analysis of its documents, leading to the proposal of writing guidelines for re-writing existing or producing new clear crisis management documents in English, which will be presented in Chapter 4.

Chapter 3 – The Crisis Management Documents and their Text Complexity

The aim of the present chapter is to describe the crisis management (CM) documents and their use, and to analyse how simple these documents are (Aluísio et al., 2008), or whether there are any high text complexity (TC) features present in them. In order to do that, the chapter will present a TC analysis of an especially collected corpus of written CM documents. Although there already are text simplification approaches for the crisis management domain (Johnson, 1993; AECMA, 1995; Renahy, et al., 2010), no text complexity analysis of English CM documents has ever been investigated. Chapter 3 is composed as follows. Section 3.1 will describe the collection, composition, and pre-processing of the corpus of crisis management documents. Section 3.2 will present the settings of the text analysis of the crisis management corpus by outlining the research hypotheses investigated, listing the analysed features and describing the further processing of the corpus. Section 3.3 will provide the corpus analysis results and their discussion, as well as criticisms of the conducted analysis. Finally, Section 3.4 will provide the conclusions.

3.1. The Crisis Management Corpus

This section will present the collected Crisis Management Corpus (CMC). Section 3.1.1 will introduce the concept of a corpus, Section 3.1.2 will describe how the CMC was collected, Section 3.1.3 will present the CMC's composition, and finally, Section 3.1.4 will present the CMC pre-processing.

3.1.1. Definitions and types of corpora

According to McEnery and Wilson (1996), a corpus is a collection of texts, collected according to specific criteria and purposes and "a basis for a form of Empirical linguistics". The word *corpus* (plural *corpora*) comes from the word for *body* in Latin. The main characteristics of a corpus are:

- Its finite size
- The fact that a corpus is a finite and limited sample of a larger population of texts
- The fact that it is in a machine-readable form
- The fact that it is a standard reference for the language in question
- The fact that it can be *annotated* (i.e. enriched with additional information, such as part-of-speech tags; phonetic, syntactic, or discourse information; anaphoric links; etc.)

Corpora can be mono- or multilingual and can contain documents from one or more domains. The CMC is monolingual and contains documents from only one domain. Because of this, it is domain-specific. This is the first collected and pre-processed English-language Crisis Management Corpus

(CMC). The next section will describe how the CMC was collected.

3.1.2. The collection of the corpus

The CMC was collected semi-automatically from the web. The sources from which the corpus was collected are various and include the U.S. Federal Emergency Management Agency (FEMA[13]), the U.S. Center for Disease Control and Prevention (CDC[14]), the World Health Organisation (WHO[15]), and the British Red Cross[16]. The corpus also contains a sub-corpus of infectious disease outbreak e-mail news downloaded from www.promedmail.org[17] and a small sub-corpus of flight operations manuals downloaded from www.smartcockpit.com[18].

The corpus was partially downloaded using the Firefox extension *Mozilla Scrapbook*[19] and partially collected manually from the above-mentioned websites. The files collected via *Mozilla Scrapbook* resulted in a collection of .html files, while the files collected manually yielded a collection of .pdf files. In order to transform the corpus to a machine-readable form, the .pdf files were converted into .txt files using the Linux command *pdftotext*, while the .html files were cleaned of html tags, formatting, and unnecessary information, and converted into raw text, via a set of Python scripts. The Python scripts which were developed differed according to the structure of the particular type of file. For example, the medical alert e-mails, downloaded from http://www.promedmail.org/ contained information which was unnecessary for the TC analysis, such as the date of publishing, reference number, headings repeated in every e-mail, source, author, and additional links. The transformation of *ProMedMail* messages to raw text was done in two steps. The first step was to

13 http://www.fema.gov/, last accessed on March 21st, 2011.
14 http://www.cdc.gov/, last accessed on March 21st, 2011.
15 www.who.int, last accessed on March 21st, 2011.
16 http://www.redcross.org.uk, last accessed on March 21st, 2011.
17 Last accessed on February 22nd, 2011.
18 Last accessed on February 22nd, 2011.
19 http://amb.vis.ne.jp/mozilla/scrapbook/, last accessed on October 21st, 2011.

remove the html tags by using the Python package *Beautiful Soup*[20]. The second step was to remove the additional information mentioned above and leave the raw text of the message. The first script was produced with the help of Raphael Rubino. *Simple English Wikipedia* articles also follow a specific formatting, called *wikimarkup,* and contain lots of additional or *meta* information, such as *infoboxes* containing a summary of the main information in the document, inter- and intra-document links, external references, external links, tools to rate, edit and disambiguate the page, etc. The *Simple English Wikipedia* articles were converted to raw text by applying Raphael Rubino's Python scripts. This was also a two-step process, consisting of a conversion from *wikimarkup* to *trectext* and then to raw text. Finally, the BNC sample corpus was converted from the original BNC annotation to raw text using a Perl script tailored to this formatting. The next section will provide more details about the composition of the CMC.

3.1.3. The composition of the corpus

The resulting Crisis Management Corpus consists of two kinds of CM documents—instructions and alerts—and of four sub-corpora:

- Instructions for general populations
- Instructions/protocols for crisis managers
- Instructions for pilots in case of emergencies
- E-mail alerts of disease outbreaks

20 http://www.crummy.com/software/BeautifulSoup/, last accessed in November 06, 2010.

Table 3.1 provides the files', sentences', and words' distributions for the whole corpus and for the separate sub-corpora. For the purposes of this thesis, the term *word-token* is defined as any occurrence of the pattern `<text>(.*[A-Za-z0-9-].*)</text>` in the text. The tags `<text></text>` refer to the output of the parser described in Section 3.1.4 and are supposed to enclose every separately found word. The amount of words has been limited according to the quantity of available files (*General Population*, *SmartCockpit*) and with a goal of not exceeding 1 million words per sub-corpus, to the extent possible.

Sub-corpus	**Number of files**	**Number of sentences**	**Number of word-tokens**
1. General Population	58 files	12 451 sentences	156 571 word-tokens
2. Specialists	160 files	74 875 sentences	1 243 381 word-tokens
3. SmartCockpit	44 files	21 511 sentences	299 175 word-tokens
4. ProMedMail	1486 files	59 477 sentences	1 029 413 word-tokens
Total for Entire CMC	1748 files	168 314 sentences	**2 728 540 word-tokens**

Table 3.1: Statistics for the entire CMC and its sub-corpora.

As can be seen in Table 3.1, the rows represent each of the sub-corpora, with the last row containing the total numbers for the whole CM corpus, while the columns contain the statistics for the individual measures. As can be seen from Table 3.1, the CMC is composed of 1748 documents of all together over 2 and a half million words, and it is characterized by a large imbalance between the numbers of files, sentences, and word-tokens of the different corpora.

The composition of the corpus is dictated by the wish to have a variety of:

- Readers:
  - General population (Sub-corpus 1 *General Population*)
  - Specialists (Sub-corpora 2, 3, and 4)
    - Crisis managers (Sub-corpus 2 *Specialists*)

    - Pilots (Sub-corpus 3 *SmartCockpit*)
    - Medical staff (Sub-corpus 4 *ProMedMail*)
- Domains and sub-languages:
    - General crisis management (Sub-corpus 1 *General Population* and Sub-corpus 2 *Specialists*)
    - Aeronautics (Sub-corpus 3 *SmartCockpit*)
    - Medical (Sub-corpus 4 *ProMedMail*)
- Document types:
    - Instructions (Sub-corpora 1, 2, and 3)
    - Alerts (Sub-corpus 4 *ProMedMail*)

The motivations for collecting the two additional sub-corpora of documents from *PromedMail* and *SmartCockpit* are to ensure variability of topics and domains. The *SmartCockpit* documents are split into different categories (flying technique, aerodynamics and performance, meteorology, navigation, engines, systems and instruments, and safety) and cover important topics such as turbulence, volcanic ash, and landing and take-off techniques in the case of stress situations. Another important reason for collecting a *SmartCockpit* sample of documents is to provide a means of analysing the aeronautics manuals, which should have been written according to the ASD-STE 100 (formal AECMA) controlled language, and determine whether they still exhibit any linguistic text complexity after the aeronautics CL has been applied. The medical alerts from *ProMedMail* (Madoff, 2004) are updated daily; they also cover a high number of topics, including human diseases, animal diseases, bioterrorism, and others, and the motivation for including them in the research is to provide a useful contribution to the biomedical sub-field of NLP.

Screenshots of an example of each kind of text (one example per sub-corpus) are provided in Figures 3.1-3.4.

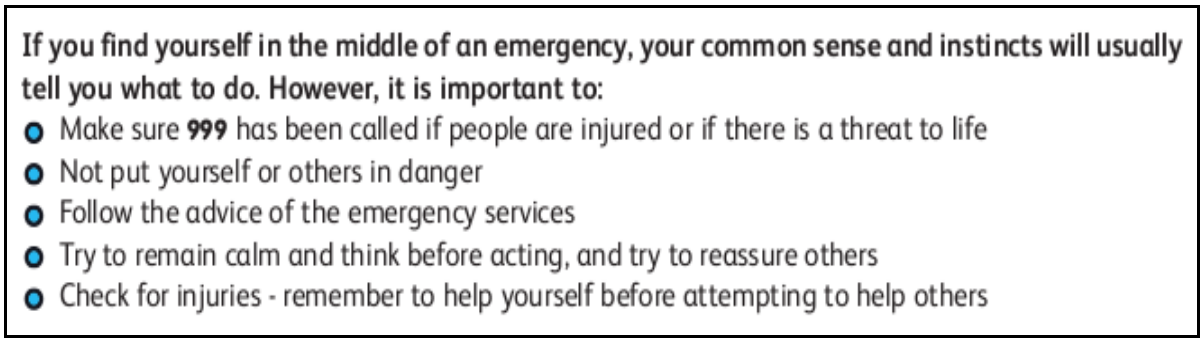


Figure 3.1: Example of Instructions for the general audience *(General Population* sub-corpus*)*.

The example in Figure 3.1 is provided from the nationally-distributed document "Preparing for Emergencies"[21], which was used as the basis of the controlled language re-writing guidelines described in Chapter 4. As can be seen, the text provides instructions for actions to be taken in an emergency situation for an audience of non-specialists.

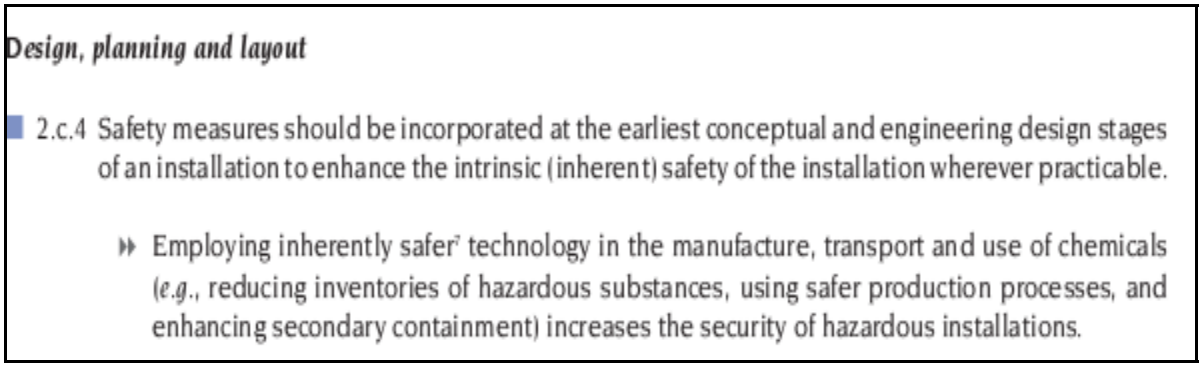
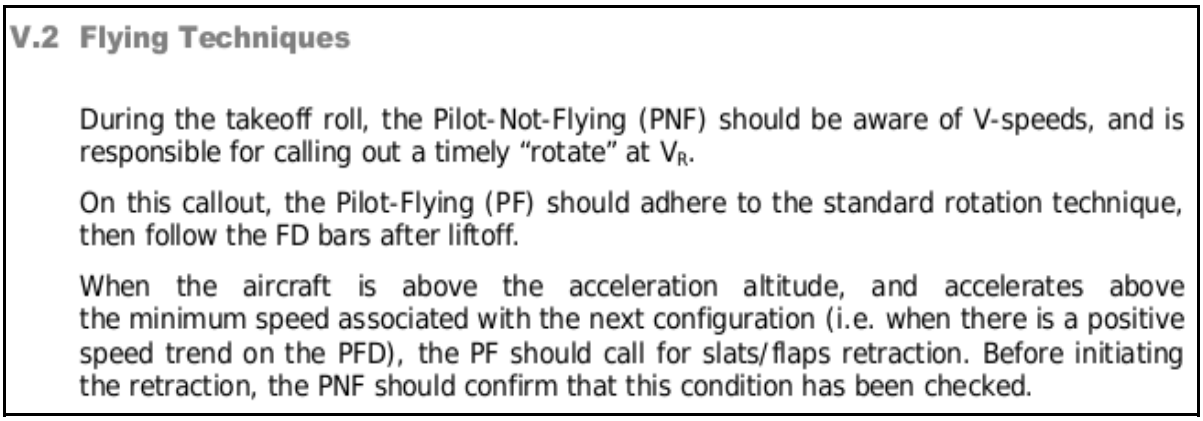


Figure 3.2: Example of plans and procedures for Crisis Managers (*Specialists* sub-corpus).

The example in Figure 3.2 is provided from a crisis management plan, which is the main type of document used by specialists in a crisis situation. This kind of document outlines the participants, their roles and responsibilities, and the procedures for each situation. As may be noted, these documents resemble legal documents.

21 http://www.direct.gov.uk/en/HomeAndCommunity/InYourHome/Dealingwithemergencies/Preparingforemergencies/DG_177092, last accessed on January 19th, 2011.



V.2 Flying Techniques

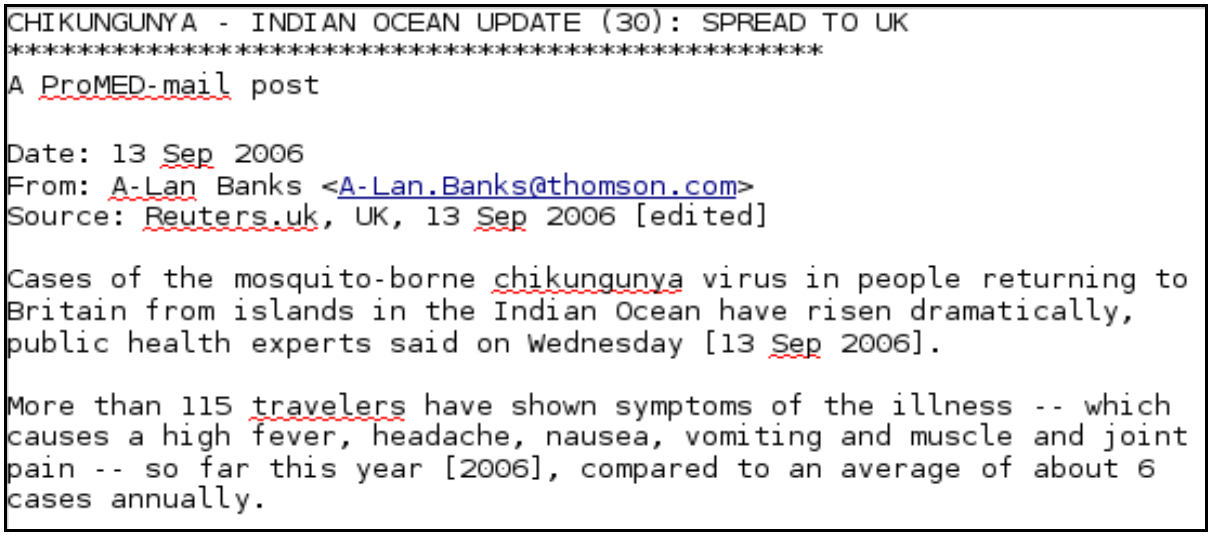
During the takeoff roll, the Pilot-Not-Flying (PNF) should be aware of V-speeds, and is responsible for calling out a timely "rotate" at V_R.

On this callout, the Pilot-Flying (PF) should adhere to the standard rotation technique, then follow the FD bars after liftoff.

When the aircraft is above the acceleration altitude, and accelerates above the minimum speed associated with the next configuration (i.e. when there is a positive speed trend on the PFD), the PF should call for slats/flaps retraction. Before initiating the retraction, the PNF should confirm that this condition has been checked.

Figure 3.3: Example of instructions for pilots (*Smart Cockpit* sub-corpus).

Figure 3.3 shows an example of instructions for flying techniques for pilots. It can be clearly noted that the instructions exhibit a particular specialist's terminology and sentence structures.



CHIKUNGUNYA - INDIAN OCEAN UPDATE (30): SPREAD TO UK
****
A ProMED-mail post

Date: 13 Sep 2006
From: A-Lan Banks <A-Lan.Banks@thomson.com>
Source: Reuters.uk, UK, 13 Sep 2006 [edited]

Cases of the mosquito-borne chikungunya virus in people returning to Britain from islands in the Indian Ocean have risen dramatically, public health experts said on Wednesday [13 Sep 2006].

More than 115 travelers have shown symptoms of the illness -- which causes a high fever, headache, nausea, vomiting and muscle and joint pain -- so far this year [2006], compared to an average of about 6 cases annually.

Figure 3.4: Example of a medical alert (*ProMedMail* sub-corpus).

Finally, Figure 3.4 provides an example of a *ProMedMail* alert. The additional unnecessary information mentioned in Section 3.1.2 is visible in the figure (*A ProMED-mail post*, *Date, From,* and *Source*). The text is not in the form of instructions, but rather in a descriptive style. Section 3.1.4 will provide details about the further preparation of the Crisis Management Corpus for the Text Complexity analysis described in Section 3.2.

3.1.4. The pre-processing of the corpus

In addition to being converted from .pdf and .html to raw text, the corpora files were further pre-processed in order to prepare them for the Text Complexity analysis. The pre-processing employed an existing state-of-the-art parser which identified single words, their parts of speech, syntactic roles in the sentence, and anaphoric links between them, as well as segmenting the individual sentences in each text. This was done in order to allow the more granular distinguishing of precise markers based on linguistic information.

The parser used for these purposes was the dependency parser *Machinese Syntax* (Tapanainen and Järvinen, 1997), which is one of the *Machinese platform* tools. The *Machinese* platform is built by *Connexor* and contains different tools, including a phrase tagger, a full syntactic parser, and a semantic role labeller.

A parser is an NLP tool which performs "parsing". Parsing, as defined by Carroll (2003), is a process that involves "using a grammar to assign a (more or less detailed) syntactic analysis to a string of words". The Connexor parser provides dependency parsing, which means that it provides additional information about the head/modifier dependency links between words. Table 3.2 lists an example of an imperative sentence taken from the CMC and its format in the *Machinese Syntax* parser output.

Sentence	**"Never open windows!"**
Parsed output	<?xml version="1.0" encoding="iso-8859-1"?> <!DOCTYPE analysis SYSTEM "http://www.connexor.com/dtds/4.0/fdg3.dtd"> <analysis><sentence id="w1"> <token id="w2"> <text>**Never**</text> <lemma>never</lemma> <tags><syntax>@ADVL %EH</syntax> <morpho>ADV</morpho></tags></token>

	<token id="w3"> <text>**open**</text> <lemma>open</lemma> <depend head="w4">attr:</depend> <tags><syntax>@A> %>N</syntax> <morpho>A ABS</morpho></tags></token> <token id="w4"> <text>**windows**</text> <lemma>window</lemma> <depend head="w1">main:</depend> <tags><syntax>@NH %NH</syntax> <morpho>N NOM PL</morpho></tags></token> <token id="w5"> <text>**!**</text> <lemma>!</lemma></token> </sentence> </analysis>

Table 3.2: An example of a parsed sentence using *Machinese Syntax*.

As can be seen from Table 3.2, the first row contains the raw text sentence “Never open windows!”, while the second row contains the parsed sentence. Since the version of *Machinese Syntax* which was employed produces XML output, the text displayed in the second row shows the original sentence enriched by XML tags, providing additional information for the whole sentence and for each word and punctuation mark in the sentence. The output starts with specifying that this is an XML document and which version of XML it is (<?xml version="1.0" encoding="iso-8859-1"?>). The second row shows that this is output of the *Connexor* parser (<!DOCTYPE analysis SYSTEM "http://www.connexor.com/dtds/4.0/fdg3.dtd">). The output of the syntactic analysis of the sentence starts with <analysis><sentence id="w1">, in which the sentence is given a reference number, and ends with the closing tags (</sentence></analysis>). Each token[22] is bracketed between the tags <token></token> and is assigned a reference number (e.g. <token id="w2">) and the following meta-information:

- Token as it appears in the text, e.g. *<text>Never</text>*
- Lemma of the token, e.g. *<lemma>never</lemma>*
- Syntactic function in the sentence, e.g. *<syntax>@ADVL %EH</syntax>*
- Part of speech, e.g. *<morpho>ADV</morpho>*

22 This thesis distinguishes between *word-tokens*, which are the single occurrences of the words, and *tokens*, which are all the instances of single words, punctuation marks and other symbols, identified by the parser as `<text></text>`.

Since it is a dependency parser, sometimes meta-information about the dependencies is provided, e.g. *<depend head="w1">main:</depend>*.

3.2. Text Complexity Analysis of the Crisis Management Corpus

The previous sections introduced the Crisis Management Corpus (CMC). This section will describe the Text Complexity (TC) analysis run over the CMC, in order to estimate the risk of high TC being a technological CM hazard as explained in Chapter 1. The TC analysis of the collected corpus is based on the assumption that text complexity can be measured by using a number of computable measurements, as has been shown already in Section 2.2, but as the existing methods are not suitable for the purposes of this thesis, a particular approach to investigating high TC has been undertaken. The approach is described in Sections 3.2.1 – 3.2.6. More concretely, Section 3.2.1 will introduce the general setting and the main concepts of the corpus analysis conducted; Section 3.2.2 will present the point of view of the high TC corpus study by listing the research hypotheses investigated, together with the analysed high TC features; Section 3.2.3 will discuss the further automatic processing of the CMC; Section 3.2.4 will list the corpus analysis results; Section 3.2.5 will present the corpus analysis findings; and finally, Section 3.2.6 will provide some criticisms of the analysis as it was run.

3.2.1. General setting of the corpus analysis

As was shown in Section 2.2, there are several old and new approaches to measuring the relative or absolute levels of TC of a text. It has been shown that the existing approaches are not suitable for

the CM domain, because the old approaches (or *readability formulae*) employ only basic markers (generally word length and sentence length) and assign a unique score corresponding to a specific student level, while the new approaches, although employing several more features, are not tailored to the CM domain. For these reasons, a domain-specific TC analysis was conducted on the corpus. The TC analysis conducted is comparative in nature and analyses the CMC in comparison with a random sample of a corpus of general English (*BNC*) and with a corpus containing simplified (using a different method) texts (*Simple English Wikipedia*).

The analysed features were of two types:

1. Text complexity features, leading to a TC analysis
2. Descriptive linguistic features, leading to a descriptive linguistic analysis

As the CMC sub-corpora are of very different natures and document types, the sub-corpora were compared to *BNC* and *Simple English Wikipedia* separately, with only the two features which are taken into account in all readability measures (word length and sentence length) being calculated for the whole CMC.

The motivations for comparing CMC with a corpus of general English consisted of the objective of making a domain language comparison with general English, while the comparison with a corpus of simplified texts was motivated by the need to have a "gold standard" for the TC comparison.

In more detail, the *BNC*, or the *British National Corpus* (Burnard, 1995), is a balanced corpus of General English, composed of 100 million words. BNC is widely used in NLP and features both

written and spoken English documents. *Simple English Wikipedia* (Simple English Wikipedia, 2009) is a version of *English Wikipedia*[23], written according to the composition rules of the Basic English controlled language (see Chapter 2 for a definition of Basic English). *Simple English Wikipedia* thus basically represents an online encyclopaedia written in simple language. There are Text Simplification approaches in NLP that employ *Simple English Wikipedia* as a comparable corpus to English Wikipedia in order to infer simplification rules, with the goal of building fully automatic TS systems (Zhu et al., 2010; Biran et al., 2011). Although most of the CMC documents contain instructions and thus should be compared with a corpus of simplified instructions, such corpora are not available; for this reason, *Simple English Wikipedia*, as the only available corpus of simplified texts, has been used. Table 3.3 provides the number of files, word-tokens and sentences for BNC and Simple English Wikipedia. For comparison with the CMC distributions, see Table 3.1.

Corpus	Number of files	Number of sentences	Number of word-tokens
BNC sample	50 files	69 212 sentences	1 401 264 word-tokens
Simple English Wikipedia	80 067 files	329 142 sentences	4 389 599 word-tokens

Table 3.3: Statistics for the *BNC* sample and *Simple English Wikipedia* corpus.

Table 3.3 is constructed in the same way as Table 3.1, i.e. the rows list the corpora and the columns give their descriptive statistics. In particular, it should be noted that the size of the BNC sample was taken in accordance with the first version of CMC, which contained only the General Population and Specialists sub-corpora and thus amounted to 1.5 million word-tokens. It was decided to keep the whole body of *Simple English Wikipedia,* as it was of a limited size. As can be seen, *Simple English Wikipedia* contains an enormous number of very short texts. Next, Section 3.2.2 will present the research hypotheses and the features which were measured in the corpus analysis.

3.2.2. Research hypotheses investigated and features analysed

23 www.wikipedia.org, last accessed on January 02nd, 2011.

This section will present the research hypotheses of the corpus analysis of CMC and the analysed features. Specifically, Section 3.2.2.1 will introduce the research hypotheses investigated and the methods chosen for testing them; Sections 3.2.2.2, 3.2.2.3, and 3.2.2.4 will present the three groups of features analysed in the corpus; and Section 3.2.2.5 will provide the motivations for the omission of some of the high text complexity features presented in Section 2.1.3.

3.2.2.1. Research hypotheses investigated

In order to investigate how simple the CM documents are (Aluísio et al., 2008) and what distinguishes the language used for communication in this domain, the comparative corpus analysis of the CM corpus has been driven by the following two fundamental research hypotheses:

Hypothesis No. 1: The crisis management documents in the corpus are too complex to be understood and need simplification.

Hypothesis No. 2: The CMC exhibits particular linguistic features making it different from general English.

Table 3.4 lists the research hypotheses together with an analysis of the metrics and characteristics that would be necessary to test them, the methods chosen to test them, and their motivations.

Hypothesis No. 1: The crisis management documents in the corpus are too complex to be understood and need simplification.	
Analysis of needs:	In order for this hypothesis to be tested, it is necessary to study the collected CMC texts for the presence of markers of high TC. In order to be able to determine whether the incidence of markers of high TC is too high or not, a comparison with the incidence of these markers in a corpus of simplified documents is necessary.
Testing Method chosen:	The CMC sub-corpora will be compared with a corpus of simplified texts (*Simple English Wikipedia*) for higher or lower incidence of high TC features.

Limitations of the method and motivations for the choice:	Due to the particularities of the CMC documents (containing mainly instructions to execute an action), the ideal case would be to compare them with a corpus of simplified documents of similar type (i.e. instructions). Due to the lack of such a corpus, *Simple English Wikipedia* will be used, since although its documents are of a different type (encyclopaedic articles and sentences which are mainly statements), it is the only available corpus of simplified texts.
Hypothesis No. 2: The CMC exhibits particular linguistic features making it different from general English.	
Analysis of needs:	In order for this hypothesis to be tested, it is necessary to study the collected CMC texts for any linguistic particularities differentiating them from general English. As a first prerequisite, a corpus of general English language is necessary, and for this reason, the *British National Corpus (BNC)* is selected. Due to the fact that *BNC* is very large, a random sample of its documents has been selected for the purpose of testing the hypothesis. Since a number of high TC features are already calculated in order to test **Hypothesis No. 1**, these features will also be used in an attempt to differentiate the CMC from the general English language. In addition, a number of purely linguistic features (proportion of nouns, adjectives, adverbs and verbs), which do not provide any decisive TC conclusions, will be assessed.
Testing Method chosen:	The CMC sub-corpora will be compared for amount of various linguistic features with a general English corpus, a random sample taken from the *BNC*.
Limitations of the method and motivations choice:	One of the limitations of the choice of this testing method is the same as for Hypothesis No. 1, as most of the documents do not contain instructions. The second limitation is that the linguistic features studied do not provide an exhaustive picture of the CM language in itself , but the motivation is that the study of the CM sub-language is beyond the scope of this thesis and for the purposes of the thesis, it is sufficient to determine whether the CM language is different from the standard one.

Table 3.4: Corpus analysis hypotheses and their testing methods.

As can be seen, Table 3.4 lists the research hypotheses, the investigations that must be carried out in order to test them, and the chosen testing methods and their limitations and motivations in consecutive rows, with those of **Hypothesis No. 1** coming first and those of **Hypothesis No. 2** coming second. As can be seen from Row 3 and Row 7 of Table 3.4, respectively, both of the methods selected to test the two hypotheses involve counting the occurrences of particular high TC and linguistic features in the texts. In particular, the following high TC and linguistic features have been examined. They have been divided into *Main high TC features (1)*, *Secondary high TC features (2),* and *Descriptive Linguistic features (3)*, but the first two groups, containing the high TC features, are also considered as being characteristic of the language under analysis, due to their linguistic nature.

The list of the text complexity issues which were studied follows below. Next to each of them is given the motivation for its use and the approach which was followed to calculate it. Due to the limitation of spreadsheets on the number of entries they can process (Griffith, 2007), series of

Python scripts have been developed especially for this purpose. For reasons of simplicity, the approach taken to calculate the numbers of these issues tried to be as shallow as possible.

3.2.2.2. *Main high TC features*

Average sentence length

High sentence length is an indicator of high syntactic complexity (see Section 2.1.3). It was selected because it is one of the most common criteria for text complexity, being used as a basis of most of the old and new readability formulae (see Section 2.2), and thus would allow comparison with other approaches and domains. It is measured as number of words per sentence. The Python script which was developed to collect this information goes through the *Machinese Syntax* parsed texts and divides the sentence lengths per number of sentences in each corpus.

Average word length

High word length is a marker of high lexical complexity (see Section 2.1.3). It was selected because it is one of the most common criteria for text complexity, being used as a basis of most of the old and new readability formulae (see see Section 2.2), and thus would allow comparison with other approaches and domains. It is measured as the number of letters per word. The Python script which was developed to collect this information goes through the *Machinese Syntax* parsed texts and divides the word lengths per number of words in each corpus.

Lexical diversity

High lexical diversity is a marker of high lexical complexity (see Section 2.1.3). It was selected because it is one of the most common criteria for text complexity, being used as a basis of most of

the new readability formulae (see Section 2.2), and thus would allow comparison with other approaches and domains. It is measured as the proportion of lemmas to inflected forms of the lemma (where lemma is the "base form" of a word and the inflected

forms are all possible forms of the word not related by derivational morphology, including the uninflected form itself; for example, for the lemma *be*, the set of inflected forms is {*be, is, are, was, were, being, been*}).

Average number of word senses

High number of word senses is a marker of high lexical complexity (see Section 2.1.3). It was selected because it is one of the most common criteria for text complexity, being used as a basis of most of the new readability formulae (see Section 2.2), and thus would allow comparison with other approaches and domains. It is measured as the number of word senses per word for any word which exists in the lexical database WordNet (described in Section 3.2.3). The Python script which was developed for this purpose goes through the *Machinese Syntax* parsed texts, extracts the words-types, then checks their presence in WordNet. If the word-type is present in WordNet, the script extracts its number of senses. The average is obtained by summing up the number of senses for all word-types found in WordNet and by dividing them per the total number of word-types which have been found to have senses in WordNet. The word-types not having senses in WordNet are not counted in this calculation.

Proportion of function words

A low number of function words indicates that there is a high number of content words (nouns, verbs, adjectives, and adverbs), which contributes to a high content word density and thus is a measure of high lexical complexity (see Section 2.1.3; Ilisei et al., 2009). This feature was selected

because it is one of the most common criteria for text complexity, being used as a basis of most of the new readability formulae (see Section 2.2), and thus would allow comparison with other approaches and domains. This thesis considers functions words to be the complement of the set of content words—the closed classes of words, such as auxiliary verbs, subordination and coordination markers, pronouns, determiners, negation markers, and prepositions. It is measured as the proportion of function words-types to the total number of word-tokens in each corpus. The Python script which was developed for this purpose goes through the *Machinese Syntax* parsed texts and calculates the proportion of function words-types from the total number of word-tokens in each corpus.

3.2.2.3. *Secondary high TC features*

Proportion of coordination markers

High number of coordination markers is a marker of high syntactic complexity. It is used in some of the readability approaches (see Section 2.2.), and is also one of the main syntactic complexity features in the Index of Syntactic Complexity score (Szmrecsanyi, 2004). It was selected because it increases sentence length and thus working memory overload, which can be dangerous in a stress situation. It is measured as the proportion of the number of "*and*", "*or*" and "*but*" to all of the word-tokens in the text. Although "," can also be a coordination marker, it has not been taken into account in this measure, as it is taken into account elsewhere while calculating the number of punctuation signs. The Python script that counts these words is case-insensitive.

Proportion of subordination markers

High number of subordination markers is a marker of high syntactic complexity. It is used in some

of the readability and text simplification approaches (see Sections 2.2 and 2.3). It was selected because it increases sentence complexity, which can be dangerous in a stress situation, since there is no time to attempt to understand all syntactic relations and dependencies under stress. It is measured as the proportion of the number of words matching the regular expressions `why` and `after|although|as|because|before|if|once|since|that|though|till|until|unless|whenever|wherever|whereas|whereupon|while|whilst` to the total number of word-tokens in the text. The Python script that calculates this measure is case-insensitive. Although some of these words are ambiguous, no further disambiguation has been done, as such disambiguation would require building a complex grammar, which is beyond the scope of this thesis.

Proportion of relative clause markers

High number of relative clause markers is an indicator of high syntactic complexity. It is addressed in most text simplification approaches (see Section 2.3), and is also one of the main syntactic complexity features in the **Index of Syntactic Complexity** score (Szmrecsanyi, 2004). It was selected because it increases sentence complexity, which can be dangerous in a stress situation, as there is no time to attempt to understand all syntactic relations and dependencies under stress. It was measured via a Python script as the proportion of the number of all word-tokens "who, when, what, where, which, why, that" to the number of all word-tokens in the parsed corpus. Although ambiguity between interrogative pronouns and relative pronouns ("*who?*"/"*who*") can arise, and the parser does provide information about relative pronouns ("*<morpho><Rel>*"), relying on the parser for this is unreliable—manual examination of the parsed text has shown that there are many errors with respect to this. For this reason, it has been assumed that these markers are most likely to be relative phrase markers if not starting with a capital letter, and thus the Python script is case-sensitive, unlike most of the other scripts discussed in this section.

Proportion of ambiguous quantifiers

High number of ambiguous quantifiers is a marker of high semantic ambiguity and thus high lexical complexity, one of the issues addressed in Graesser et al. (2006), because it is considered to increase semantic ambiguity and decrease precision, which this thesis considers to be important in a crisis situation. It was measured via a Python script as the proportion of the number of all word-tokens "*some, many, most, few, any, little, much, less, fewer, more, someone, somebody, anybody, anyone*" to the number of all the word-tokens in the corpus. The script is case-insensitive.

Proportion of punctuation signs

High number of punctuation signs is a marker of high syntactic complexity (Aluísio et al., 2008). It is considered to increase both sentence length and sentence complexity and to therefore be dangerous in a stress situation, as it can cause working memory overload and delay in the execution of instructions. It was measured via a Python script as the proportion of punctuation signs "*,,:(.-*" to the full count of tokens (all tokens, including words, punctuation signs and other symbols) occurring in the parsed text.

Proportion of discourse markers

High proportion of discourse markers is an indicator of high text cohesion, and thus low TC (Aluísio et al., 2008; McNamara, 2010). It was calculated using a Python script as the proportion of the word-tokens *First |Second |Third |Next |thus |therefore |firstly |Firstly |Secondly |secondly |then |moreover |however |finally* (which through concordance analysis were discovered to be the only discourse markers present in CM texts) to the total number of word-tokens in the text. The ambiguity of "*first*", "*second*", "*third*" and "*next*" between their use as discourse markers and their use as adjectives was resolved by specifying in the script that they must start with a capital letter

and be followed by a comma. Some examples are shown below:

- Example 1[24]:

"***First,*** *the building should be located farther back from the edge of the fill closest to the flooding surface.* ***Second,*** *the higher the basement floor is elevated, the less the risk.*"

- Example 2[25]:

"*Learn* ***First*** *Aid Kit*"

"*The* ***second*** *zone covers a broader area.*"

As can be seen from Examples 1 and 2, the words "*first*" and "*second*" in the two sentences presented in Example 1 do constitute discourse markers, and are clearly delimited by a capital letter and a comma, while the same words in the two sentences in Example 2 are not discourse markers but adjectives.

Proportion of personal and possessive pronouns

High proportion of personal and possessive pronouns is a marker of high ambiguity of anaphoric reference, and thus a high syntactic or discourse complexity. It is one of the issues most commonly addressed in automatic text simplification and in detecting text complexity (see Chapter 2). It was measured via a Python script as the proportion of the number of personal subject (*it, he*, *they*), genitive (possessive, *your, theirs, mine*) and accusative word-tokens pronouns (*them, it*), as indicated by the parser, to the total number of all word-tokens in each corpus. The script was case-insensitive.

24 Source file *1_10-01* from *Specialists* sub-corpus.
25 Source file *areyouready_full* from *General population* sub-corpus.

3.2.2.4. *Descriptive Linguistic features*

Proportion of nouns (tokens) (Aluísio et al., 2008)

This measure was selected because besides being an interesting linguistic feature, it also enters into the Index of Syntactic Complexity as a marker of high syntactic TC. It was measured via a Python script as the proportion of the number of nouns-tokens to the total number of word-tokens in the parsed text for each corpus.

Proportion of verbs (tokens)

This measure was selected because besides being an interesting linguistic feature, it also enters into the Index of Syntactic Complexity as a marker of high syntactic TC. It was measured via a Python script as the proportion of the number of verbs-tokens to the total number of word-tokens in the parsed text.

Proportion of adjectives (tokens)

This measure was selected because besides being an interesting linguistic feature, it can also be indicative of the TC of the text, because a high number of adjectives contributes to a high number of elements to remember and thus to working memory overload. It was measured via a Python script as the proportion of the number of adjectives-tokens to the total number of word-tokens in the parsed text.

Proportion of adverbs (tokens)

This measure was selected because besides being an interesting linguistic feature, it can also be

indicative of the TC of the text, because a high number of adverb modifiers contributes to a high number of elements to remember and thus to working memory overload. It was measured via a Python script as the proportion of the number of adverb-tokens to the total number of word-tokens in the parsed text.

3.2.2.5. Features not analysed in this study

As can be seen, not all of the high TC issues described in Section 2.1.3 have been analysed in this corpus analysis. The list of issues which are not investigated and the reasons why these measures have not been calculated follows below:

- **Number of Technical terms**: there is no available domain lexicon.
- **Words with high age-of-acquisition**: the texts are not for children.
- **Abstract concepts**: it is considered that only concrete concepts would be used in documents of type "instructions."
- **Words with large orthographic neighbourhood**: this high TC issue only affects readers suffering from dyslexia, while this thesis attempts to address the average reader.
- **Inconsistent terminology**: there is no available domain lexicon.
- **Figurative language**: as for abstract concepts, it is considered that figurative language will not be used in CM documents; additionally, are no reliable resources or methods to evaluate its presence in text yet (see Chapter 2).
- **Number of syntactic relations**: the method of calculating it would employ the parser's

markers, which as has already been demonstrated (Siddharthan, 2003) is unreliable and can introduce many calculation errors.

- **Passive voice**: as has been explained in Chapter 2, it is only possible to calculate the number of passive voice markers "*-ed*" followed by "*by*", which as has been demonstrated, underestimates the actual number of passives in text, sine it does not detect agentless passives or sequences of passives (e.g. *burned or otherwise damaged by*) (Cohen et al., 2010).

- **Negative constructions**: as was explained in Chapter 2, in the current state of NLP it is extremely difficult to automatically identify negation and its scope in text, and a special domain-specific grammar would be necessary, which is not available at the moment.

- **Illogical order of statements**: as was explained in Chapter 2, in the current state of NLP, automatically identifying illogical order of statements is impossible.

Next, Section 3.2.3 will present the automatic processing of the corpus, leading to the corpus analysis results, which will be presented in Section 3.3.

3.2.3. Further processing of the corpus

As was explained in Sections 3.1.2 and 3.1.4, in order to transform the corpus into a machine-readable format and prepare it for the corpus analysis, the documents were first converted from .pdf and .html format to raw text and in then parsed with Connexor's *Machinese Syntax* parser. The corpus analysis then proceeded by running a set of specially developed Python scripts over the texts. The main functioning of these scripts was described in Section 3.2.2. They take as input the parsed texts and provide as output numerical descriptions of features. However, in order to obtain

all of the results, and specifically the number of word senses, the employment of two other tools and resources was necessary: NLTK and WordNet. NLTK, or the Natural Language Toolkit (Bird et al., 2009), is a set of Python modules which allow effective NLP processing techniques. These techniques range from accessing existing corpora or importing one's own corpus, to building and graphing frequency distributions of different linguistic phenomena, creating or accessing existing lexical resources, obtaining text from the Web, performing text pre-processing (such as tokenizing, lemmatizing, and POS-tagging), classifying texts, extracting information from text, syntactic parsing, and discourse processing, to XML and HTML processing. Since NLTK is based on the programming language Python, which was used for the rest of the corpus analysis, it was also the natural choice for calculating the number of word senses per word.

WordNet (Fellbaum, 1998) is an electronic lexical database which can be accessed through NLTK. In WordNet, English words are organized into synonym sets, called "synsets", each one representing a lexical concept. WordNet was used in order to extract the number of senses per word.

The next section will present the corpus analysis results obtained on the basis of the analysis methods outlined in Section 3.2 and with the assistance of the programming tools described in the current section.

3.3. Corpus analysis results, findings and criticisms

This Section will present the crisis management corpus analysis results (Section 3.3.1), findings (Section 3.3.2) and criticisms (Section 3.3.3).

3.3.1. Corpus analysis results

As stated in Section 3.2.2.1, this study aims to test the following research hypotheses:

Hypothesis № 1: The crisis management documents in the corpus are too complex to be understood and need simplification.

Hypothesis № 2: The CMC exhibits particular linguistic features making it different from general English.

In order to test the first research hypothesis, the CMC sub-corpora will be compared with a corpus of simplified texts (*Simple English Wikipedia*) for higher or lower incidence of high TC features. The method for testing the second research hypothesis will consist of comparing the amounts of various linguistic features in the CMC sub-corpora with a random sample of the general English corpus *BNC*.

This section presents the results obtained by running the set of Python scripts on the CMC, the Simple English Wikipedia, and the random sample of BNC in order to test the research hypotheses, presented in Section 3.2.2.1 and to measure the presence of the features presented in Sections 3.2.2.2, 3.2.2.3 and 3.2.2.4.

The descriptive statistical results of the three groups of features (*Main high TC features*, *Secondary high TC features* and *Descriptive Linguistic features*) are provided in Tables 3.5, 3.7, and 3.8, respectively. As a reminder, the first two groups of features (*Main high TC features* and *Secondary high TC features*) are used for testing **Hypothesis № 1**, the TC comparison of the CMC with *Simple English Wikipedia*, while all of the three groups are used for testing **Hypothesis № 2**, i.e. the differences between the CMC language and general English (*BNC*). At the end of the section, Table

3.9 provides some cumulative results for the CMC as a whole. The results were obtained by developing Python scripts that process the outputs of the corpus analysis scripts (described in Section 3.2.2) as input files and calculate means, standard deviation, standard error of the mean, and statistical significances at 95% and 99% confidence levels.

The reported results (mean ± standard error of the mean) are rounded to the third digit after the decimal point and are significant at 99% confidence level. Table 3.5, Table 3.7 and Table 3.8 have the same structure. The analysed features are presented in the vertical columns, while the horizontal rows list the different corpora. For the purposes of the two corpus analysis investigations, the results must be compared along the vertical axis. Table 3.5 lists the *Main high Text Complexity features* (estimated means for sentence and word lengths, lexical diversity, average number of word senses, and the proportion of function words to the total number of word-tokens).

Corpus/Features	Average sentence length (in words)	Average word length (in letters)	Lexical diversity (types/word-tokens ratio)	Average number of word senses	Proportion of function words (function words/word-tokens ratio)
1. General Population	12.575 ± 0.275	5.114 ± 0.020	0.042 ± 0.001	8.275 ± 0.061	0.390 ± 0.003
2. Specialists	16.606 ± 0.160	5.709 ± 0.008	0.017 ± 0.0002	7.082 ± 0.018	0.335 ± 0.001
3. SmartCockpit	13.908 ± 0.268	5.222 ± 0.014	0.027 ± 0.001	7.857 ± 0.041	**0.341 ± 0.002**
4. ProMedMail	17.307 ± 0.131	5.285 ± 0.009	0.025 ± 0.0004	7.235 ± 0.021	**0.343 ± 0.001**
Simple English Wikipedia	13.336 ± 0.043	4.764 ± 0.003	**0.022 ± 0.0001**	8.026 ± 0.012	0.383 ± 0.0006
BNC Sample	20.246 ± 0.133	4.923 ± 0.006	**0.023 ± 0.0003**	8.110 ± 0.021	0.424 ± 0.001

Table 3.5: Results for the Main high TC features for the six corpora.

The groups of values in bold represent that the means in this group have likely overlapping values and that there is little or no difference between them. There are two groups of such values in Table 3.5 and they are:

- *Smart Cockpit* and *ProMedMail* for Function words
- *Simple English Wikipedia* and *BNC Sample* for Lexical diversity

The fact that none of the CMC sub-corpora exhibit high similarity with the *BNC Sample* confirms the second hypothesis, that there are linguistic differences between the CM language and general English.

Regarding the highest and lowest values, as is clear from Column 2, the highest average sentence lengths are found in sentences from the *BNC Sample*, while the shortest are found in *Simple English Wikipedia*. This result supports Hypothesis No. 1, that the CMC sub-corpora are more complex than

a corpus of simplified English, and also supports Hypothesis №. 2, that the corpus of general English differs from the CMC. The much lower average word length in BNC than in the CMC sub-corpora also supports the second hypothesis. Otherwise, comparably large differences were not found between the word length values, with the highest word lengths being found in the *Specialists* sub-corpus, and the lowest found in *Simple English Wikipedia*, but the word lengths are significantly different, which again supports the first hypothesis, as it shows that the word lengths in the CMC sub-corpora are larger than those of the corpus of simplified texts. The lexical diversity values are more or less at the same level, except for *General Population,* where they are the highest, which supports the first hypothesis, and points to the fact that this CMC sub-corpus desperately needs lexical diversity reduction. A similar situation is noted with respect to the average number of word senses. On the one hand sub-corpora 2, 3, and 4 have similar values regarding this TC issue and are lower than *Simple English Wikipedia* and the *BNC Sample* (the latest confirming the second research hypothesis regarding specialist corpora). On the other hand, they are the highest for *General Population*, and higher than *Simple English Wikipedia*, which again confirms the first research hypothesis and points out that *General Population* words desperately need ambiguity reduction. The values for the proportion of function words are lower for sub-corpora 2, 3, and 4 than for *Simple English Wikipedia*, and higher for *General Population,* with the highest being *BNC Sample*. These results show that sub-corpora 2, 3, and 4 suffer from high content word density and thus support the first hypothesis. Also, the fact that the *BNC Sample* value is much higher than the CMC sub-corpora supports the second hypothesis. As for calculating the number of word senses per word, the number of lemmas (types) from each corpus which are present in WordNet is important, such numbers are provided in Table 3.6.

Corpus/Features	**Number of unique lemmas**	**Lemmas in WordNet**	**Lemmas NOT in WordNet**
1. General Population	6 605	5 888	717
2. Specialists	20 800	13 457	7 343

3. SmartCockpit	8 148	6 385	1 763
4. ProMedMail	26 478	14 321	12 157
Simple English Wikipedia	95 081	34 422	60 659
BNC Sample	32 132	23 803	8 329

Table 3.6: Numbers of total unique lemmas and those in WordNet and not per corpus.

In Table 3.6, as in Table 3.5, the rows display the data for each of the corpora. However, the columns in Table 3.6 provide the number of unique, not repeating lemmas as total, present in WordNet, and not present in WordNet. Table 3.6 shows the striking difference between *Simple English Wikipedia* and the other corpora, consisting in the fact that almost two-third of the lemmas in *Simple English Wikipedia* are not present in WordNet. Next, Table 3.7 presents the results for all of the corpora for the *Secondary high TC features*, namely:

- Proportion of coordination markers
- Proportion of subordination markers
- Proportion of relative clause markers
- Proportion of ambiguous quantifiers
- Proportion of punctuation signs
- Proportion of discourse markers
- Proportion of personal and possessive pronouns

Corpus/Features	**Prop. of Coor. mark./word-tokens ratio**	**Prop. of Sub. mark./word-tokens ratio**	**Prop. of Rel mark./word-tokens ratio**	**Prop. of Ambig. quant./word-tokens ratio**	**Prop. of Punct. mark./all tokens ratio**	**Prop. of Disc. Mark./word-tokens ratio**	**Prop. of Pron./word-tokens ratio**
1. General Population	0.048 ± 0.001	0.025 ± 0.001	0.014 ± 0.0008	**0.009 ± 0.001**	**0.115 ± 0.002**	**0.0015 ± 0.0002**	0.042 ± 0.001

2. Specialists	0.052 ± 0.000	**0.0169 ± 0.0002**	**0.0116 ± 0.0002**	0.005 ± 0.0001	0.104 ± 0.0001	0.0012 ± 0.00001	**0.009 ± 0.0002**
3. SmartCockpit	**0.0323 ± 0.0008**	0.019 ± 0.001	**0.0123 ± 0.0005**	0.006 ± 0.0003	0.097 ± 0.001	0.0021 ± 0.0002	**0.009 ± 0.0004**
4. ProMedMail	0.033 ± 0.000	**0.0174 ± 0.0003**	0.0126 ± 0.0002	0.007 ± 0.0002	0.111 ± 0.001	**0.0014 ± 0.00001**	0.014 ± 0.0003
Simple English Wikipedia	**0.0322 ± 0.0002**	0.020 ± 0.0001	0.016 ± 0.0001	**0.009 ± 0.0001**	**0.115 ± 0.0003**	0.0018 ± 0.00005	0.034 ± 0.0002
BNC Sample	0.039 ± 0.000	0.023 ± 0.000	0.019 ± 0.0003	0.010 ± 0.0002	0.096 ± 0.001	0.0032 ± 0.0001	0.039 ± 0.0004

Table 3.7: Results for the Secondary high TC features for the six corpora.

In Table 3.7 the highest values for coordination markers are those of CMC sub-corpora 1 and 2, with all of the CMC corpora being higher than the values of *Simple English Wikipedia*, which supports the first hypothesis. Regarding the subordination markers, all CMC sub-corpora have lower values than *Simple English Wikipedia*, except for sub-corpus 1 (*General Population*), which has a higher value and thus supports the first hypothesis. For relative markers, *Simple English Wikipedia* and the *BNC Sample* have the highest values, which supports the second hypothesis, that the language of the CMC corpus of texts is different from that of general English. The ambiguous quantifier values are similar for all of the corpora, with higher values being found in *General Population* and *Simple English Wikipedia* and the *BNC Sample,* which has the highest values, which again supports the second hypothesis. Proportions of punctuation markers exhibit large differences between the single corpora, with *General Population* and *Simple English Wikipedia* being the highest and those of the *BNC Sample* and the *SmartCockpit* being the lowest, which can support the second hypothesis. As for discourse markers, the three of the four CMC sub-corpora have lower values than *Simple English Wikipedia,* which supports the first hypothesis and shows that these sub-corpora need to be more coherent, while the amount of the discourse markers in the *BNC Sample* are much higher than those of the CMC sub-corpora, which supports the second hypothesis, that general English is different from the CM domain language. There are also large differences between the proportion of pronouns between the different corpora, with CMC sub-corpora 2, 3, and 4 having lower values than those of *Simple English Wikipedia*, but *General Population* having a much higher

number, which supports the first hypothesis. Also, the difference in the amount of pronouns of the CMC sub-corpora 2, 3, and 4 from the *BNC Sample* supports the second hypothesis, at least for corpora for specialists.

As can be seen, similarly to Table 3.5, Table 3.7 also has groups of corpora which have likely overlapping values for some of the features (listed in bold). The similar values in Table 3.7 are:

- *SmartCockpit* and *Simple English Wikipedia* for Coordination markers
- *Specialists* and *ProMedMail* for Subordination markers
- *Specialists* and *SmartCockpit* for Relative markers and Pronouns
- *General Population* and *Simple English Wikipedia* for Ambiguous quantifiers and Punctuation markers
- *General Population* and *ProMedMail* for Discourse markers

Again, the fact that none of the CMC sub-corpora exhibit such high similarity with the *BNC Sample* confirms the second hypothesis, that there are linguistic differences between the CM language and general English.

Although even if some of them have no difference at the 99% level but have some small differences at the 95% level, this thesis considers such small differences as being irrelevant. For this reason, the pairs in bold are considered to have the same value. Examples of such changes relative to the level of statistical significance follow below.

There is no statistical significance at 99% nor at 95% between the Subordination markers ratios for

Specialists and *ProMedMail*. There is also no statistical significance at the 99% level between the Coordination markers ratios for *SmartCockpit* and *Simple English Wikipedia*, but there is at 95%: the value for *SmartCockpit* is 0.0323 ± 0.0006, while the value for *Simple English Wikipedia* is 0.0322 ± 0.0001. As can be seen, the differences are too small to be considered. The same occurs with the differences between the *Specialists* and *SmartCockpit* for Relative markers–there is no difference at the 99% level, but there is a difference at the 95% level: the value for *Specialists* is 0.0116 ± 0.0001, while the value for *SmartCockpit* is 0.0123 ± 0.0004.

Note that since there is a very low quantity of discourse markers per corpus, their values are provided listing up to the fifth sign after the decimal point. Table 3.8 lists the remainder of the analysed features, namely the *Descriptive Linguistic features* (noun, verb, adjective, and adverb proportions from the total word-tokens numbers per corpus). The purpose of this data is to cast light on the linguistic characteristics of the six corpora, but some of them (nouns and verbs) can also be used for calculating the **Index of Syntactic Complexity** (Szmrecsanyi, 2004), already presented in Chapter 2.

Corpus/Features	**Proportion of Nouns (nouns/word-tokens ratio)**	**Proportion of Verbs (verbs/ word-tokens ratio)**	**Proportion of Adjectives (adjectives/word-tokens ratio)**	**Proportion of Adverbs (adverbs/word-tokens ratio)**
1. General Population	0.360 ± 0.003	0.141 ± 0.002	**0.084 ± 0.002**	**0.045 ± 0.001**
2. Specialists	0.423 ± 0.001	**0.099 ± 0.001**	0.099 ± 0.001	0.026 ± 0.0004
3. SmartCockpit	0.444 ± 0.002	**0.099 ± 0.001**	0.080 ± 0.001	0.034 ± 0.001
4. ProMedMail	**0.388 ± 0.001**	**0.099 ± 0.001**	**0.084 ± 0.001**	0.039 ± 0.0005
Simple English Wikipedia	**0.388 ± 0.001**	0.118 ± 0.0003	0.072 ± 0.0003	**0.042 ± 0.0002**
BNC Sample	0.305 ± 0.001	0.120 ± 0.001	**0.086 ± 0.001**	0.056 ± 0.0005

Table 3.8: Results for the Descriptive linguistic features for the six corpora.

As can be seen from Table 3.8, the highest values for the proportion of nouns are for *Specialists* and *SmartCockpit*, with the value of *BNC Sample* being the lowest, which supports the second

hypothesis. For verbs, the highest value is found in *General Population*. The lowest values for verbs are those of sub-corpora 2, 3, and 4, which have identical values and large differences with the *BNC Sample,* which again supports the second hypothesis as related to specialists' corpora. The highest proportion of adjectives is found in the *Specialists* sub-corpus, while the lowest is found in *Simple English Wikipedia.* Finally, there is also a large difference between the proportions of adverbs, with the highest value being that of *BNC Sample* and the lowest, that of *Specialists*. The much higher value for adverbs in the *BNC Sample* supports the second hypothesis.

Among the *Descriptive linguistic features* there are also groups of similar values. They are listed in bold, while the non-overlapping values have a default font. The overlapping values are spread across all the features and are:

- *ProMedMail* and *Simple English Wikipedia* for Nouns
- *Specialists*, *SmartCockpit,* and *ProMedMail* for Verbs
- *General Population, ProMedMail,* and *BNC Sample* for Adjectives
- *General Population* and *Simple English Wikipedia* for Adverbs

The similarity of the proportion of adjectives between sub-corpora 1 and 4 and the corpus of general English is the only factor we have observed which is not consistent with Hypothesis No. 2. Finally, Table 3.9 presents the total cumulative results for the CMC as a whole. Due to the large differences between the genres of the individual CM sub-corpora, the results for the CM corpus as a total have been calculated only for word length and sentence length, which as has been seen in Chapter 2, are the two most frequently calculated features in all of the state-of-the-art approaches to Measuring Text Complexity. The total means for the whole CM Corpus are provided in the first row of Table

3.9. The last two rows contain the mean estimations for *Simple English Wikipedia* and the *BNC Sample*, already presented in Table 3.5. The means are given with their standard errors.

Corpus/Features	Average sentence length (in words)	Average word length (in letters)
CMC Total	16.211± 0.098	5.462 ± 0.005
Simple English Wikipedia	13.336 ± 0.043	4.764 ± 0.003
BNC Sample	20.246 ± 0.133	4.923 ± 0.006

Table 3.9: Totals for the whole CM corpus.

Table 3.9 shows the totals for word and sentence length for the whole CM corpus, compared with the values already presented in Table 3.5 for the *BNC Sample* and *Simple English Wikipedia*. The results are again significant at the 99% level, due to the large number of entries. As can be seen, the total value of the average sentence length for the whole CMC stands between those of the *BNC Sample* and *Simple English Wikipedia*, with the value of the CMC corpus being higher than the value of the corpus of simplified English, which supports the first hypothesis. Furthermore, the substantial difference with the value of the *BNC Sample* supports the second hypothesis. Although there are again no large differences for the values of average word length, the total value for the whole CMC is clearly higher than those of the *BNC Sample* and *Simple English Wikipedia,* which supports both the first and the second hypotheses in the sense that on the one hand there is a clear difference between the CMC and the corpus of general English, and on the other hand, that the CMC presents a higher complexity than the corpus of simplified English. Next, Section 3.3.2 will provide discussion of the above presented results and draw findings from them.

3.3.2. Corpus analysis findings

This section will lay out the findings from the corpus analysis as related to the research hypotheses presented in Section 3.2.2.1 and the obtained results, listed in Section 3.3. As a reminder, the

investigated research hypotheses, along with the methods for testing them, were the following:

Research hypothesis No. 1: The crisis management documents are too complex to be understood and need simplification. Method to test: Compare the numbers of *Main* and *Secondary TC features* (Tables 3.5 and 3.7 in Section 3.3) of the CMC sub-corpora with the numbers of the same features in *Simple English Wikipedia*.

Hypothesis No. 2: The CMC exhibits particular linguistic features, making it different from general English. Method to test: Compare the numbers of all of the three groups of features (Tables 3.5, 3.7 and 3.8 in Section 3.3) of the CMC sub-corpora with the numbers of the same features in the *BNC Sample*. For the purposes of the research findings, *Simple English Wikipedia* will be referred to as the *simplified corpus*, while the *BNC Sample* will be referred to as the *reference corpus*. The analysis of the findings related to each of the two hypotheses follows below.

3.3.2.1. Research hypothesis No. 1 findings

As has been seen in Table 3.9 in Section 3.3, comparison of the two primary main TC features characterizing the CMC as a whole (word length and sentence length) shows that CMC has much higher values, and thus greater TC, than *Simple English Wikipedia* (sentence length CMC 16.211± 0.098 vs. 13.336 ± 0.043; word length CMC 5.462 ± 0.005 vs. 4.764 ± 0.003). In detail, as Tables 3.5 and 3.7 have shown, among the four sub-corpora, *ProMedMail* exhibits the highest sentence length, followed by *Specialists*, *SmartCockpit,* and *General Population*, while all of the sub-corpora have higher word length values than *Simple English Wikipedia* (simplified corpus).

In relation to the other three *Main high TC features*, only *Specialists* has a lower lexical diversity

than the simplified corpus, while all of the others have higher values, especially *General population,* which has a very high value. From this it can be concluded that the *General Population* documents badly need the construction of a CM lexicon, and adherence to it.

Although *ProMedMail*, *Specialists*, and *SmartCockpit* present a lower number of word senses per word than the simplified corpus, *General Population* again presents a higher value, indicating that its word meanings need to be restricted and lexical ambiguity needs to be reduced.

In relation to the function/content words ratio, the three last sub-corpora present lower values of function words than the simplified corpus, which means that their lexical density is very high and needs reduction. This could be explained by the fact that they are all characterized by specialised terminology.

The *Secondary high TC features* show that *Specialists* and *General Population* exhibit a higher proportion of coordination markers, and thus higher syntactic complexity, than *Simple English Wikipedia*. The same finding holds for the subordination markers value of *General Population*, which indicates that the texts in this corpus exhibit a high sentence complexity that needs to be addressed. The highest numbers (higher than *Simple English Wikipedia*) of relative clause markers are present in *SmartCockpit* and *ProMedMail*, which indicates that there is a high incidence of relative clauses in these specialised corpora, which have to be split into simple sentences. The proportion of ambiguous quantifiers in the three specialised corpora is at lower levels than those in the simplified corpus, which has a similar value to *General Population*. Although the presence of such markers is not an issue in a free encyclopaedia, it is of vital importance for the CM instructions and thus needs to be addressed. *General Population* again has the highest value for punctuation signs, similar to the simplified corpus, which needs to be further investigated, as it has already been

seen that *Simple English Wikipedia* has shorter sentences. The proportion of discourse markers in *General Population, Specialists,* and *ProMedMail* is lower than in *Simple English Wikipedia,* which means that the texts of these corpora need to be made more cohesive. Finally, the proportion of personal and possessive pronouns in the corpora show that the three specialised corpora need less anaphora resolution than the *General Population* sub-corpus—one which needs pronoun replacement badly.

On the basis of the aforementioned findings, it can be concluded that:

- All of the CMC sub-corpora exhibit TC issues that need to be addressed.
- Most of the CMC sub-corpora exhibit higher values for the main TC issues than the corpus of simplified texts, which supports the first hypothesis.
- The separate CMC sub-corpora are characterized by different sets of high TC issues.
- Some of the CMC sub-corpora exhibit similar values for some of the high TC issues.
- There are similarities between the amounts of TC issues present in some of the sub-corpora (the values listed in bold in Tables 3.5 and 3.7).

The instructions for the general population present sufficient complexity issues to motivate its choice as the first type of text for controlled language development. In particular, the analysis has shown that the high TC issues of *General Population* which most need simplification are a specific set and that it is most necessary to:

1. **Decrease**:
    1. Word length

2. Lexical diversity
3. The number of word senses
4. The proportion of coordination markers
5. The proportion of subordination markers
6. The proportion of ambiguous quantifiers
7. The proportion of punctuation markers
8. The proportion of personal and possessive pronouns

2. **Increase** the proportion of discourse markers

3.3.2.2. Research hypothesis № 2 findings

In testing this hypothesis—that the CMC has particular linguistic features that make it different from general English—the primary features on which the findings are based are the Descriptive linguistic features in Table 3.8. In particular, all of the four sub-corpora show a higher proportion of nouns, and a lower proportion of verbs, than *BNC Sample*. This could mean that the CMC sub-corpora are more action-orientated, which could be explained by the fact that most of them contain instructions. In contrast, in terms of adjectives, *Specialists* has a much higher number of them than the reference corpus, which means that it is much more descriptive. The values of *General Population* and *ProMedMail* are similar to *BNC*. The amount of adverbs also shows differences, with the three specialised corpora having a lower number of adverbs than the reference corpus, while *General Population* has a higher value than the other three corpora, but still lower than the *BNC*.

In addition, Tables 3.5, 3.7, and 3.9 present clear differences between the CMC sub-corpora, the

corpus as a whole, and the *BNC Sample*. In particular, Table 3.9 shows that the sentences of CMC are much shorter than those of the reference corpus, while the words of CMC are longer than those of the reference corpus. Table 3.5 also shows differences at the levels of proportion of function words, which are much more frequent in the reference corpus than in the CMC sub-corpora, while Table 3.7 shows that among the Secondary high TC features, there are differences between the CMC sub-corpora and *BNC Sample* in relation to the proportion of ambiguous quantifiers and discourse markers, as well as pronouns (which are higher in *BNC Sample*). This could be explained by the fact that *BNC Sample* is probably more explicative and has more cohesive text than the CMC. Another observation is the fact that the values of *General Population* for certain markers (number of word senses; number of subordination markers; number of relative markers; proportion of personal and possessive pronouns; and proportion of nouns, verbs, adjectives and adverbs) are similar to those of *BNC Sample*; this could be motivated by the fact that the *General Population* language is close to General English.

The findings show that indeed there are language differences and TC differences between General English and the CM English, and for this reason a CM-language-specific TS approach must be developed and the existing state-of-the-art approaches are not applicable in this case. Next, Section 3.3.3 will present the criticisms of the this analysis.

3.3.3. Criticisms of the conducted analysis

As has been seen in several points of the presentation of the chosen corpus analysis methodology and the discussion of the obtained results, the selected approach, although designed and conducted as well as possible, has some limitations. The next three sections will present its limitations and provide some suggestions for their solutions.

3.3.3.1. Criticisms of the choice of indicative high TC issues

Although the state-of-the-art has provided valid proofs that the high TC issues chosen for the corpus analysis are the most appropriate ones, some of them contradict each other. The cases that contradict each other are listed below:

- Word length vs. number of word senses
- Number of punctuation marks

The first issue contradiction arises from the facts that although short words are considered to be more comprehensible than long ones, as shorter words have higher frequency, according to the Zipf's law (Zipf, 1949), they also usually have a higher number of word senses. The second issue is also controversial, as it could also be interpreted in two ways: more punctuation could be a sign of higher text complexity, but more punctuation could make sentence structure more explicit and thus clearer. A further, more detailed analysis of these issues is required.

3.3.3.2. Criticisms of the methodology of detecting high TC markers

In addition to the difficulties already mentioned in Section 3.2.2 which were encountered while automatically processing the CMC sub-corpora for detecting the relevant high TC markers (such as the ambiguity of subordination markers, the ambiguity of relative clause markers, the ambiguity of discourse markers, etc.), another issue was discovered. It is described below:

Ambiguity of adjectives

The analysis of the identified adjectives has shown that the parser commits many errors in marking words as adjectives. Examples taken from the *General Population* sub-corpus are:

Example 1: "*For **further** information*". "*Further*" is an adjective, but is labelled by the parser as a determiner. The parser output is *<text>further</text> <morpho>DET</morpho>*.

Example 2: "*Specifications for **netting** materials*". "*netting*" is an adjective, but in the parser output *<text>netting</text>* is labelled as a noun: *<morpho>N NOM SG</morpho>*.

Example 3: The opposite type of mistake is observed, I.e. when the word is not an adjective, but it is recognized as such. The verb "*put*" is recognized as an adjective. For the input "***Put** on protective gloves.*", the parser output is: *<text>Put</text> <lemma>put</lemma> <morpho>A ABS</morpho></tags>*.

3.3.3.3. Criticisms of the choice of linguistic resources

As has been mentioned in Section 3.3 about 2/3 of the lemmas in each corpus are present in WordNet. Table 3.6 showed the respective numbers of lemmas in and not in WordNet for each corpus. A further analysis has shown that the lemmas not existing in WordNet are:

- The most frequent function words in English language (such as "*the*", "*and*", "*for*", "*of*", and "*that*")
- Specialised terminology

- Abbreviations ("*vhs*", "*fifa*")
- Foreign names and words ("*millennio*", "*chiquita*")
- Spelling errors ("*singificant*", "*diferent*")
- Lemmatisation errors ("*Chiva*" from "*Chivas*")

In order to overcome this limitation of WordNet, the WordNet resource could be enriched with additional concepts and their relationships, with the use, for example, of an English dictionary (Nastase and Szpakowicz, 2003) or of a domain-specific ontology (Poprat and et al., 2008), although no such ontology exists for the crisis management domain, and therefore that technique is not applicable here. Next, Section 3.4 will provide the conclusions of this chapter.

3.4. Conclusions

This Chapter has studied how simple the crisis management (CM) texts are and thus the levels of presence of high text complexity features in the text corpora collected especially for this purpose. A comparison with a corpus of simplified texts (Simple English Wikipedia) has been conducted in order to determine whether the level of TC of the Crisis Management corpus (CMC) is higher than is desirable and whether text simplification is necessary. In addition, a comparison with a corpus of General English has been conducted, in order to assess whether the CM language is different from General English.

In order to test the two hypotheses, sets of Python scripts were developed in order to clean, convert,

and prepare the corpus data, and to conduct the corpus analysis. Additionally, linguistic resources (BNC and WordNet) and programming tools (NLTK) were used. The results obtained on the basis of the aforementioned analysis demonstrate that the Crisis Management documents indeed exhibit a higher number of TC issues than the simplified texts, and thus text simplification is necessary. They also demonstrate that the CM language appears to be different than the General English language, and thus a specialised text simplification approach, tailored to this domain, needs to be developed. The results also show a worrying number of high text complexity issues affecting the *Instructions for General Population*. Next, Chapter 4 will present the manual controlled language text simplification approach which has been developed specifically for the *Instructions for General Population.*

Chapter 4 – The Controlled Language for Crisis Management

Computers are useless. They can only give answers. (*Pablo Picasso)*

This chapter will present the Controlled Language for Crisis Management (CLCM) – adapted specifically for simplifying emergency instructions written in English. The structure of the chapter is as follows: Section 4.1 is the Introduction of the chapter, Section 4.2 will present the context in which CLCM has been developed. Section 4.3 will present the CLCM guidelines and rules and discuss the high TC issues that they address, and will also introduce the CLCM prototype for Bulgarian which has been developed. Section 4.4 will compare CLCM with the existing controlled languages, listed in Chapter 2, then with its source controlled language LiSe, and finally will describe it according to an existing CL specification framework. The Chapter will be closed with a Conclusions Section (Section 4.5) introducing the motivations for an extensive evaluation of CLCM, described in Chapters 5, 6, and 7.

4.1. Introduction and Motivations

As explained in Section 1.2, the second aim of this thesis is to propose a method for re-writing crisis management documents into clear and straightforward ones and of creating clear and straightforward CM from scratch.

Chapter 3 showed that, although the simplicity and comprehensibility of texts used in the crisis management domain are crucial, they exhibit a high number of high text complexity issues which would hinder human comprehension. More concretely, the TC corpus analysis showed that the document type *Instructions for the General Population* (IGP) exhibits a high number of high TC issues. This is very dangerous because unlike specialists or pilots, the general population has not been trained on interpreting these documents, so their comprehension of instructions in emergency situations is crucial.

As has been discussed in Chapter 2, the natural solution to high text complexity is Text Simplification (TS). The findings of the analysis in Chapter 3 and specifically in Section 3.3.2.1, indicate that there is a case for applying it to instructions for the general population in order to simplify the already existing ones, or in order to produce new ones simple enough to be read. Although a variety of TS approaches exist, ranging from manual to semi-automatic and fully automatic ones (as reviewed in Chapter 2), the existing approaches are not suitable for this particular task. The semi-automatic and automatic approaches to TS address only a limited number of high TC issues, and most of them cause some information loss as redundant text elements are deleted. As stated in the text simplification definition of this thesis, the aim is to simplify (Section 2.3.1) without any information being lost. This is particularly important for instructions for the general population, as losing any information may cause dangerous consequences, and

simplification should be done in such a way as to address as many high TC issues as possible, in order to avoid any hindrance to human comprehension. These two reasons make the semi-automatic and fully automatic approaches inappropriate for simplifying emergency instructions. For this reason, the manual approaches, i.e. the controlled languages, as presented in Chapter 2, are considered more appropriate, since they can be designed more easily to address a large number of high TC issues and to preserve information. The existing CLs are, however, not suitable for the CM domain, because they do not reflect the document structure and the sublanguage characteristics of emergency instructions. For this reason, a specific controlled language for such documents and language should be used. However, the only CLs existing for the crisis management domain are the French CL LiSe (Renahy et al., 2010, developed for *Protocols* in the health-care and the CM domain) and PoliceSPEAK (Johnson, 1993), restricted to communication of officials involved in the management of the Channel Tunnel. The controlled language described in this chapter has been adapted from the controlled language LiSe existing for French. In order to ensure improvement in text comprehension of the *Instructions for the General Population* (IGP), the CLCM rules have been checked for compliance with psycholinguistic findings about comprehension under stress (Kiwan et al., 1999; Ogrizek et al., 1999) and with respect to the high text complexity issues which hinder comprehension (Harley, 2008). The list of such issues is presented in Section 2.1.

4.2. The Context of CLCM

This section presents the context in which the Controlled Language for Crisis Management has been developed. Namely, Section 4.2.1 present the MESSAGE project, while Section 4.2.2 will discuss the emergency instructions characteristics and high text complexity issues, which CLCM addresses.

4.2.1. The MESSAGE Project

The first prototype of CLCM was originally developed in the context of the MESSAGE Project[26], an EU-funded project, which started on December 31st, 2007 and had a duration of twenty months. It was intended to address the problem that communication during the management of an emergency situation is the weakest link, because natural language exhibits a high number of complexity and ambiguity issues at different levels, which can hinder human comprehension. The Project aimed at providing a solution to this problem by developing a methodology for reducing the complexity and ambiguity of the language used for communication in emergency situations. In particular, the project aim was to develop controlled language guidelines containing rules imposing restrictions on the allowed lexical units, syntactic structures, and general presentation of the information for writing messages, protocols and alerts for situations arising from terrorism and other security-related risks. In this way, the guidelines aimed to promote writing standards for documents used in such situations. The development of standards was conducted in close collaboration with a variety of end-users who would apply the CL guidelines to writing their documents. Some of the partners are: Sandwell Council Resilience Unit (United Kingdom), Autoroutes-Trafic (France), French Air Force, Airbus (France), the journal Geomedia (Bulgaria)[27], National Police (Spain), Fire-fighters (Greece), Centre for Veterinary Inspection (Poland), etc. As can be seen from the variety of end-users, the CL technology was finally designed to be applied not only to security-related risks, but also to various man-made and natural emergencies and crises. The CL language methodology was first developed for French by the project's coordinators (Centre Tesnière, Université de Franche-Comté, Besançon, France), who already have more than twenty years of experience in controlled languages, and then transferred to the national languages of the

26 Full title: Alert Messages and Protocols, project financed by the European Union (JLS/2007/CIPS/022). With the support of the Prevention, Preparedness and Consequence Management of Terrorism and other Security-related Risks Programme European Commission - Directorate-General Justice, Freedom and Security.

27 http://www.geomedia.bg/, last accessed on April 5th, 2012.

project's partners (Research Group in Computational Linguistics, University of Wolverhampton, Wolverhampton, UK; Departament de Filologia Francesa i Romànica, Universidad Autònoma de Barcelona, Spain; and Instytut Romanistyki, Uniwersytet Warszawski, Poland). The Project produced several deliverables. Its main contribution was the set of CLs for emergency situations for four EU languages (French, English, Spanish and Polish). Additionally, the Project produced a specially designed training course for end-users willing to apply the CL rules to their documents[28]; a kit facilitating the transfer of the methodology to new EU languages or domains[29]; a network of linguists and end-users specially trained in applying this methodology; and an international conference with published proceedings featuring participants from both the linguistic and NLP communities and crisis management end-users[30]. More information about the MESSAGE project can be found on the coordinators' project's website[31], while an advertising leaflet and slides from the training course for end-users are available at the website of the British partner[32]. In this context, the British partner has developed guidelines for writing emergency documents which were later adapted to emergency instructions for the general population, as this was considered to be the most critical type of communication. A description of the documents follows below.

28 The UK version of it can be accessed at: http://clg.wlv.ac.uk/projects/Message/, last accessed on January 18th, 2011.
29 http://message-project.univ-fcomte.fr/resources-en.html, last accessed on January 18th, 2011.
30 www.ismtcl.org, last accessed on January 18th, 2011.
31 http://message-project.univ-fcomte.fr/, last accessed on January 18th, 2011.
32 http://clg.wlv.ac.uk/projects/Message/ (last accessed on January 18th, 2011).

4.2.2. Textual analysis of *Instructions for the General Population*

In order to determine whether CLCM can be adapted from the French CL LiSe, a manual analysis of the information content, structure, and language comprehensibility of relevant documents was conducted in order to investigate whether the English emergency documents, and particularly the *Instructions for the General Population*, exhibit the same characteristics as the French documents for which LiSe was developed. The analysis was run on a document[33] provided by the British end-user, the Sandwell Council Resilience Unit, which was disseminated nation-wide, and some emergency preparedness documents downloaded from the Web (which were included in the Crisis Management Corpus, already described in Chapter 3). A screenshot of the document (*Preparing for Emergencies)* can be seen in Figures 4.1 and 4.2.

Figure 4.1: The title page of Preparing for Emergencies

33 http://www.direct.gov.uk/en/HomeAndCommunity/InYourHome/Dealingwithemergencies/Preparingforemergencies/DG_177092, last accessed on January 19th, 2011.

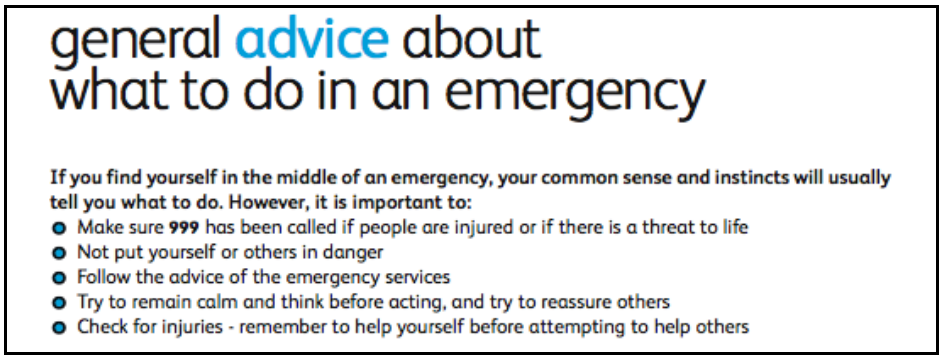


Figure 4.2: An example of a specific topic from Preparing for Emergencies

The other two documents analysed together with *Preparing for Emergencies* are *Are You Ready? An In-depth Guide to Citizen Preparedness*, a comprehensive guide for emergency situations published by the American Federal Emergency Management Agency (FEMA) and *Individual Preparedness and Response to Chemical, Radiological, Nuclear, and Biological Terrorist Attacks* (Davis et al., 2003). The information content, structure, and language of these documents were manually analysed, and CLCM was tailored to them. The aim of these documents is to provide instructions for the general audience to follow in an emergency situation in order to preserve their health, lives, property, or children. The examined documents usually contain an introductory section explaining the target readers, aims, contents, and publishing body of the guide, as well as a table of contents divided into sections providing general and concrete instructions for emergency situations. The documents try to distinguish between separate situations and to provide instructions for a specific safe behaviour for every one of them. The documents happen to provide additional information, such as quotes from famous people or citations of other related documents. All of the English documents are written using the same or similar terms, but the terms are usually not previously introduced and defined.

According to this analysis, although the emergency instructions do not have an unique structure, they always contain all or some of these textual elements:

- title of the document
- titles of the subsections
- condition or an emergency situation definition
- instructions for actions to follow in a specific situation or under a certain condition
- comments, such as explanations, definitions, citations, or warnings
- lists of items
- pictures
- illustrative graphics

The coordinators of the MESSAGE project (Centre Tesnière, Besançon, France) have performed an extensive manual corpus analysis and have identified a number of specific text comprehension problems existing in emergency instruction documents (Renahy, 2009; Renahy et al., 2010; Renahy et al., 2011). Although a comprehensive automatic TC analysis of the whole Crisis Management corpus has already been presented in Chapter 3 and has shown the exact frequencies of each of these issues, a manual and more detailed analysis of the aforementioned documents was done in order to quickly analyse whether some of the issues cited by Renahy et al. (2009) which are difficult to detect automatically are also present in the English-language texts. Even though the examined documents have been specifically prepared for non-specialist readers and have the goal of making it easy to identify the important information and the order of actions, the manual analysis tested and proved the hypothesis that they still exhibit high text complexity issues. Some examples which can hinder the presentation of information and readers' comprehension follow. These include:

- unclear titles or situations which are not clearly distinguished in the text
- logical or chronological contradictions between instructions
- unimportant information showed more clearly than important information
- syntactic reading difficulties
- lexical reading difficulties

Some of these issues are discussed and illustrated by examples in the sections below. Examples of different lexical terms used to denote the same concept have been already shown in Section 2.1.3.2.

4.2.2.1. Unclear titles or situations which are not clearly distinguished in the text

Since it is important to know which actions to take in which situations, the title of the whole document and of each sub-section must be clearly visible, easily identifiable, and as readable as possible.

Unfortunately, very often the emergency documents do not identify clearly the titles of their sections, which makes it difficult to understand to which situation the listed instructions are referring. An example is provided on page 6 of *Preparing for Emergencies* and can be seen in Figure 4.3.

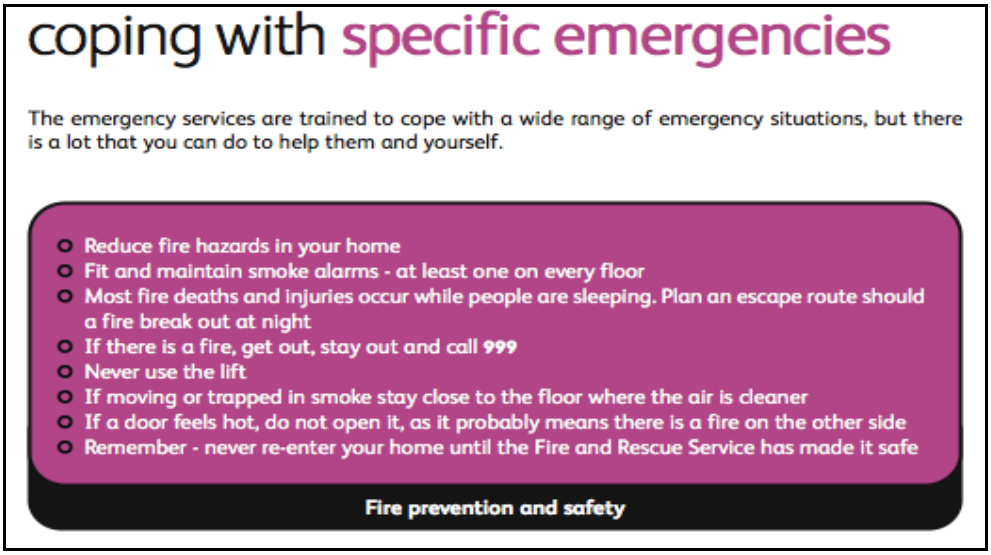


Figure 4.3: Example of an unclear title.

As can be seen in Figure 4.3, the main title of the page seems to be *Coping with specific emergencies*. An explanation and several instructions follow. However, while reading the instructions, the reader begins to have the impression that all of them are about fire safety, until getting to the last instruction and discovering that *Coping with specific emergencies* has been the title of the whole section, while the actual title of these instructions is *Fire prevention and safety*, which is printed at the end of the section. This is an example of a title, which is not clearly stated and of a situation which is not clearly distinguished in the text.

4.2.2.2. Logical or chronological contradictions

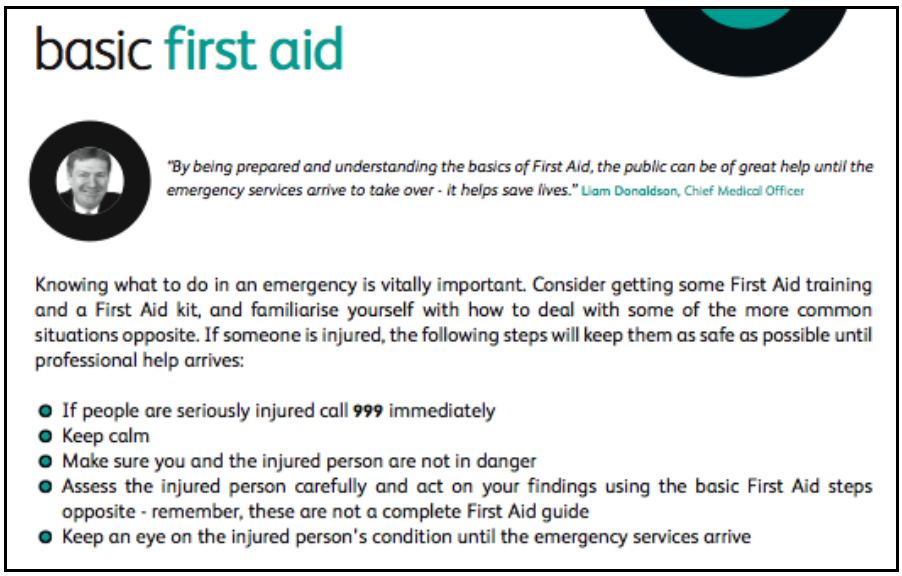
Logical or chronological contradictions between instructions are also an issue. Very often, in addition to the difficulty in identifying the title of the situation or identifying the situation itself, there are also difficulties created by the illogical order of instructions. The most frequent case of a logical/chronological contradiction is when two actions which usually follow one after another are listed in the instructions with the second action provided before the first one. An example can be seen in the same page of *Preparing for Emergencies* as the previous example in Figure 4.4.

Figure 4.4: Example of an illogical order of instructions.

As can be seen in Figure 4.4, the instructions are not given in logical order. In particular, the instruction *Never use the lift* is provided after *If there is a fire, get out and call 999*, while a person very often uses a lift in order to get out of a building, and thus the instruction not to use a lift should be given before the one about getting out.

4.2.2.3 Unimportant information shown more clearly than important information

Unimportant information being shown more clearly than important information is also an issue. As there is no technical standard for writing emergency instruction documents, very often the documents are structured as a narrative book, which makes it difficult to discover the important information (e.g. situation definitions and instructions) in the text. An example is taken again from *Preparing for Emergencies* and is provided in Figure 4.5.


basic first aid

"By being prepared and understanding the basics of First Aid, the public can be of great help until the emergency services arrive to take over - it helps save lives." Liam Donaldson, Chief Medical Officer

Knowing what to do in an emergency is vitally important. Consider getting some First Aid training and a First Aid kit, and familiarise yourself with how to deal with some of the more common situations opposite. If someone is injured, the following steps will keep them as safe as possible until professional help arrives:

- If people are seriously injured call **999** immediately
- Keep calm
- Make sure you and the injured person are not in danger
- Assess the injured person carefully and act on your findings using the basic First Aid steps opposite - remember, these are not a complete First Aid guide
- Keep an eye on the injured person's condition until the emergency services arrive

Figure 4.5: Example of unimportant information.

The main aim of *Page 12*, shown in Figure 4.5, is to provide instructions for basic first aid, as can be understood from the main title of the page. Although it is logical that a definition of the cases when such aid would be needed and concrete instructions should follow the title, a quote from a medical specialist is listed first. This would distract the attention of the reader and hinder finding the important information when this booklet is consulted during an emergency situation.

4.2.2.4 Syntactic reading difficulties

In addition to the issues concerning the document structure or the order of presentation of information, very often there are cases of well-known syntactic issues which can hinder readers' comprehension, such as long sentences, the use of the passive voice, etc. The use of the passive voice is shown in the first instruction listed in Figure 4.2 – *Make sure 999 is called...* Overly long sentences make comprehension problematic, as people may not have the time to read and process them all during an emergency situation, or the ability to remember all the information and its order after reading it. An example is provided in Figure 4.4. It shows an instruction in which the action and the condition are hidden behind a long explanation:

"*Most fire deaths and injuries occur while people are sleeping*. *(Explanation)* ***Plan an escape route*** *(Action) should a fire break out at night. (Condition)*"

Since the results of the analysis showed that the *Instructions for the General Population* for English have very similar content and comprehensibility problems to the *Protocols for specialists* writing in French for which LiSe was developed, CLCM was created by adapting the LiSe guidelines for *Protocols for specialists* to the specificities of the *Instructions for the General Population* in English. The next section will describe the CLCM guidelines, which have been created on the basis of the LiSe guidelines and tailored to the TC issues in the *Instructions for the General Population*.

4.3. The CLCM guidelines

This section will present the CLCM guidelines. Section 4.3.1 will present the structure of the guidelines, Section 4.3.2 will discuss the types of rules, Section 4.3.3 will indicate which high text complexity issues the CLCM rules address. Finally, Section 4.3.4 will describe the experiment of adapting CLCM to Bulgarian.

4.3.1. Presentation of the guidelines

The CLCM guidelines are provided in Appendix B of the present thesis. They are inserted in a booklet-type document with a precise structure and consist of thirty-six pages listing over eighty rules. The document starts with the table of contents, which introduces the structure of the guidelines. As the aim of MESSAGE project was CLs which would respect the writing standards established in the project (Cardey et al., 2010), the structure of the CLCM guidelines (provided in

Figure 4.6) follows that of the LiSe guidelines for *Protocols for Specialists*.

- General settings
- Grammatical terms mini-dictionary
- General rules valid for the whole document
- Guidelines for step-by-step document writing
- Sets of rules relative to each specific document sub-part
- The allowed syntactic structures
- The forbidden syntactic structures
- The lexical rules
- The forbidden lexical expressions
- A domain dictionary
- A step-by-step re-writing example
- A re-written emergency instructions example

Figure 4.6: CLCM guidelines structure.

As can be seen from Figure 4.6, the structure is flexible enough to allow introducing controlled language rules and specifications for different sub-languages and different domains, while also allowing the insertion of other useful resources for helping to follow the rules and write easy-to-understand documents.

4.3.1.1. CLCM *General Settings* section

The CLCM "General Settings" Section, similarly to the first section of the LiSe guidelines, provides general information about the types of rules and explanations of their notation. As in LiSe, every CLCM rule is characterized by a unique reference number which is composed of letters and numbers indicating which type of document and which part of the document the rule refers to, together with a consecutive number. An example is "*PrDurS_T_S_3*", which is a rule referring to protocols ("*Pr*") to follow during an emergency ("*Dur*"), designed for writers who are specialists in the domain ("*S*"); in addition the rule is restricting the writing of the document title ("*T*") and it is the 3rd ("*3*") syntactic rule ("*S*").

4.3.1.2. CLCM grammatical terms mini-dictionary

The next section, the "Grammatical terms mini-dictionary", is provided in order to address end-users' lack of technical competence in linguistic terminology. It lists the grammatical terms used in the guidelines, together with a lay definition. This section has not been specified completely, in order to allow flexibility to potential end-users. An example taken from it is shown in Table 4.1.

Term	Definition	Examples
Part of Speech	Part of speech refers to the terms by which we categorise words.	Noun, verb, adjective.

Table 4.1: Example of a term from the mini-dictionary of grammatical terms

As it can be seen in Table 4.1, each term (column 1) is provided with a lay definition for writers who are not linguists (column 2) and with examples (column 3) which can be found further on in the guidelines.

The proper controlled language rules are listed in the following three sections ("**General rules valid for the whole document**", "**Guidelines for step-by-step document writing**", and "**Sets of rules relative to each specific document sub-part**").

4.3.1.3. CLCM general rules valid for the whole document

The "**General rules valid for the whole document**" introduce the document by specifying its aims and composition (the list of compulsory and optional document sub-parts and their relative order).

4.3.1.4 CLCM guidelines for step-by-step document writing

The “Guidelines for step-by-step document writing” of CLCM correspond to LiSe's “list of general rules for the whole document”. They guide the writer through writing the whole document, helping him to organise information, introducing the document sub-parts one by one, and mentioning all formatting restrictions that divide them. This section also contains the rules which are applicable everywhere in the document and are not sub-part-specific (e.g. specific only to writing the title). Examples of such rules are listed below:

- “In_F_03: Separate each block of instructions with a new line.”
- “In_G_13: Write only one piece of information per line.”
- “In_L_01: Choose the words in accordance with the lexical rules on p. 28.”
- “In_L_05: Keep preposition and verb together in phrasal verbs.”
- “In_P_01: If you make reference to a specific document: Put the document title in quotation marks.”

4.3.1.5. CLCM sets of rules for specific document sub-parts

The last of these three main sections, “Sets of rules relative to each specific document sub-part”, lists the concrete rules for each specific document sub-part, exactly as in LiSe. As was mentioned before, a manual analysis of some emergency instruction documents has shown that their structure is very similar to the structure of *Protocols for specialists*, for which the LiSe guidelines were designed. For this reason, and in order to ensure uniformity and stricter structure of the *Instructions*

for the General Population, CLCM defines a set of document sub-parts very similar to LiSe: instructions (providing the actions needed to be undertaken), conditions (providing the situations in which the unique actions have to be undertaken), comments (providing additional information with lower importance), lists of items, titles, and titles of sub-sections. More information about the document sub-parts defined by CLCM will be provided in Section 4.3.2, while the concrete differences between the CLCM and LiSe sets of document sub-parts will be discussed in Section 4.4.2. A few examples of rules from these concrete sets follow below.

Guidelines for writing a title:

"In_T_G_01: Write a title that describes unequivocally only this document."

Guidelines for writing a title of a section or a sub-section:

"In_St_F_01: Use the following formatting:

Font style: **bold**.

Font size: at least one unit bigger than text and one unit smaller than title.

Alignment: left< . "

Guidelines for writing a comment:

"In_Cm_P_01: If the comment is a warning:

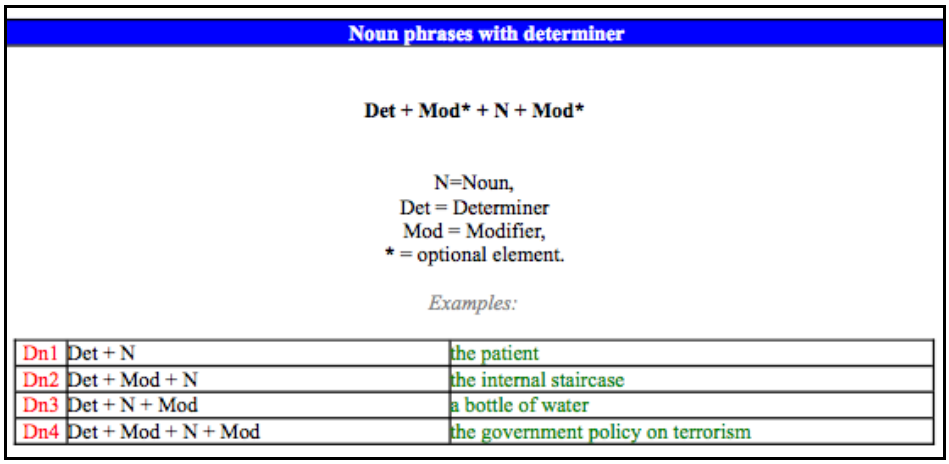
Put an exclamation mark at the end of the comment.

If not:

Put a dot at the end of each comment. "

4.3.1.6. CLCM allowed syntactic structures

The section “Allowed syntactic structures”, similarly to the corresponding section in LiSe, defines the syntactic structures allowed for use in the sub-parts of the document by specifying their elements and their relative order. This section also provides examples of sentences which are a realisation of these syntactic structures. The CLCM's set of syntactic structures and their concrete realisations have been adapted for English from the French syntactic structures in the LiSe guidelines. There are five syntactic structures currently listed: conditional clause, imperative clause, alphanumeric sequence, noun phrase (NP) without determiner, and NP with determiner. For each type of syntactic structure, the allowed combinations, together with example sentences, are listed. An example of an allowed syntactic structure is shown in Figure 4.7.



Noun phrases with determiner

Det + Mod* + N + Mod*

N=Noun,
Det = Determiner
Mod = Modifier,
* = optional element.

Examples:

Dn1	Det + N	the patient
Dn2	Det + Mod + N	the internal staircase
Dn3	Det + N + Mod	a bottle of water
Dn4	Det + Mod + N + Mod	the government policy on terrorism

Figure 4.7: Example of an allowed syntactic structure.

Figure 4.7 shows the syntactic structure “Noun phrases with a determiner”, which is used in titles or as an element of other syntactic structures. Examples in English are provided.

4.3.1.7. CLCM forbidden syntactic structures

“Forbidden syntactic structures” is a section which allows the end-user or the linguist to add

syntactic structures which have to be avoided in order to not risk reducing instructions' comprehensibility. Only a few forbidden syntactic structures are currently present in the Guidelines. An example are "garden path sentences" (Harley, 2008), which can be defined as sentences whose syntactic structure leads to the expectation of a different sentential meaning from the intended one. An example of such a sentence is "*The old man the boat.*", which misleads the reader due to the part-of-speech ambiguity of the words "*old*" (adjective/noun) and "*man*" (noun/verb). Due to these ambiguities, the first impression of the reader is that "*The old man*" is the subject of the sentence, and the expectation is that the next word would be a verb. Instead the sentence continues with the surprising determiner "*the*", which makes the reader go back to the beginning of the sentence in order to try to figure out a different parse. This return to re-parse the sentence would result in a longer processing time to extract the sentence information. This could be dangerous in an emergency situation, because in such situations the reader is expected to react as fast as possible.

4.3.1.8. CLCM lexical rules and forbidden lexical expressions

The "Lexical rules" section allows listing general lexical rules which apply to the whole document, such as "*Use only literal meaning*" (as has been explained in Chapter 2, both abstract concepts and figurative language can create comprehension problems for readers because the reader first has to access the literal meaning of the expression and then has to try to figure out the non-literal one). Next, the "Forbidden lexical expressions" section allows the user to list forbidden lexical expressions or classes of lexical items, such as ambiguous words, pronouns, technical terms, etc. (whose comprehensibility difficulty has been also explained in Chapter 2). These last two sections are also in a prototype version, and for this reason contain only a few entries.

4.3.1.9. CLCM domain dictionary

The next section, "Domain dictionary", allows adding lists of alternative paraphrases for technical terms in different domains. This section is needed in order to help a writer who is not a domain specialist by providing him/her with the means to understand the relevant technical terms. The current prototype features only one domain dictionary ("First Aid Medical Terminology"). An extract from it is provided in Figure 4.8.

TFAMT_0002	ambulance	Ambulances (Pl.)	A vehicle used to transport sick or injured people with medical needs. Ambulances can be cars, trucks, helicopters, boats, or airplanes. **Also Known As:** Mobile intensive care unit (MICU), rescue units, medical transport units.	Post emergency telephone numbers by phones (fire, police, **ambulance**, etc.).	
TFAMT_0003	amnesia		A condition in which memory is disturbed and/or lost		Loss of memory Source: Plain English Campaign

Figure 4.8: Example from the domain dictionary

As can be seen from Figure 4.8, the first column lists the term's reference number "*TFAMT_numbers*" (with TFAMT standing for "Term First Aid Medical Term" and the numbers indicating the term number); the second column provides the term itself (in this case "*ambulance*" and "*amnesia*"); the third column lists the alternative grammatical forms of the term (a plural form in the case of a noun ("*Ambulances, Pl.*") or an adjective and a person and tense in the case of a verb); the fourth column provides a lay-level definition of the term, together with other known synonyms; the fifth column can contain examples of sentences in which the current term is used; finally, the last column can suggest a term with which the current one could be replaced, which is usually easy to understand by lay readers.

4.3.1.10. CLCM step-by-step examples and rewritten texts

Finally, the last two sections provide examples of how the text should look when it is rewritten

according to the guidelines. In particular, the "Step-by-step rewriting example" section shows the writer how to rewrite a complex text (provided before the rewriting example) into a simple one, sentence-by-sentence. The original text whose step-by-step transformation is presented is an extract taken from the manual *The power of humanity*, Module 6 – *How to save a life*, addressed to 10- to 14-year-old readers and available at http://www.redcross.org.uk/powerofhumanity[34]. The purpose of the module is to teach basic first aid skills for an emergency situation. The whole module is designed for informal education, and thus it is supposed to be written in a sufficiently readable way for the intended readers. A screenshot of the instructions from which the extract is taken is provided in Figure 4.9.

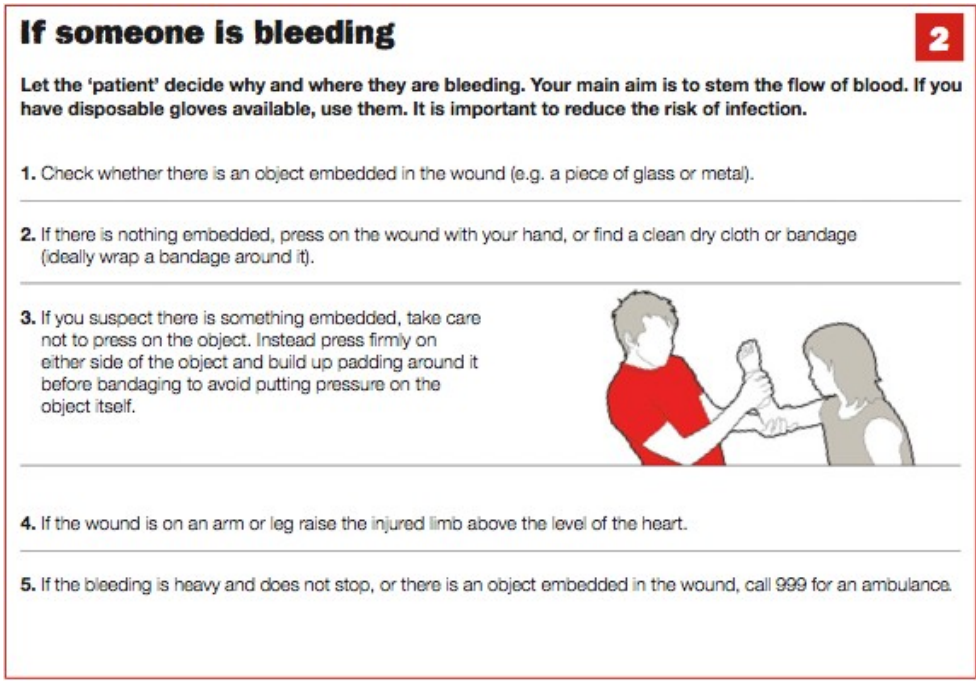


Figure 4.9: Red Cross first-aid instructions for 10-14-year-old readers.

As can be seen in Figure 4.9, the extract represents the third instruction of the text. It seems apparent that this is the instruction which provides the most information, and it is not presented in a manner which is easy to follow. A picture is provided whose aim is probably to explicate the text of the instruction, but the picture is not clear and actually corresponds to the subsequent instruction. The text of instruction 3 exhibits many high text complexity issues, such as:

34 Last accessed November 25th, 2011.

- Main information (conditions and actions) is not clearly distinguished from additional information (explanations), nor are relations of dependency between them, or their relative order:

- "*If you suspect*" (condition) … "*take care*" (action) … "*Instead press*" (action) … "*and build up*" (action) … "*bandaging*" (action) … "*to avoid*" (explanation)".

- It is not clear whether the two actions "*press*" and "*build up*" should be run in parallel or one after another, because the coordinating conjunction "*and*" is ambiguous and can allow either interpretation.

- It is not clear which action "*to avoid putting pressure on the object itself*" is an explanation of.

- Long sentences (the second sentence, "*Instead press firmly on either side of the object and build up padding around it before bandaging to avoid putting pressure on the object itself.*") are present.

- Pronouns are used ("*build up padding around IT*").

- Negation is present ("*take care NOT to press*").

- Syntactic ambiguity is present ("*and*").

- Inconsistent terminology is used ("*something*"/"*the object*").

- Omissions are present ("*bandaging .. what?*").

The result of applying the CLCM simplification rules to this instruction in order to solve these high text complexity issues is shown sentence-by-sentence in the Step-by-Step rewriting example. Figure

4.10 shows an extract of it.

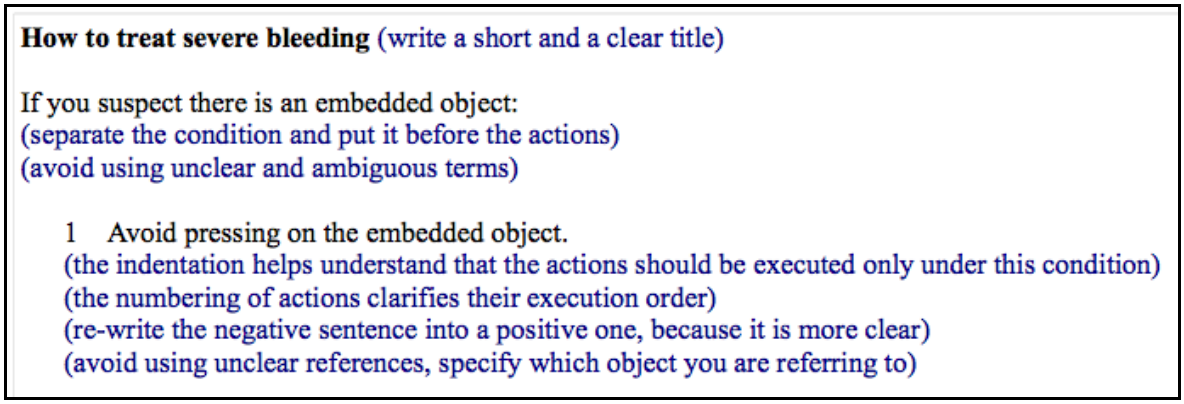


Figure 4.10: Extract from the Step-by-Step re-writing example.

As can be seen in Figure 4.10, the guiding example transforms the sentences one by one and adds explanations of what has been done and why this needs to be performed after each operation. No rule numbers or definitions are provided, in order to help the writer's comprehension of the simplification process.

The last section is the same as the one in the LiSe guidelines. It is called “A Re-written Example”. It pairs examples of the text already examined in the previous section “before” being rewritten and “after” being rewritten. This view helps the writer to see the improvement of the comprehensibility of the instruction. Table 4.2 shows the exact example provided in the Guidelines.

Before	**After**
If you suspect there is something embedded, take care not to press on the object. Instead press firmly on either side of the object and build up padding around it before bandaging to avoid putting pressure on the object itself. /Passage, taken from www.redcross.org “How To treat severe bleeding”/	How to treat severe bleeding If you suspect there is an embedded object: 1. Avoid pressing on the embedded object. 2. Do the following actions simultaneously: - Press firmly on either side of the embedded object. - Build up padding around the embedded object. *Explanation: This needs to be done in order to*

	avoid putting pressure on the object itself. 3. Finally, bandage the wound.

Table 4.2: Before and After Section of the CLCM.

As can be seen from Table 4.2, the first column (entitled "Before") contains the original (complex) text, while the second column ("After") contains the text rewritten in a more visible and clearer fashion, with the following changes being made:

- The instruction is composed of short sentences.
- The instruction is provided with a sub-section title ("*How to treat severe bleeding*").
- The condition ("*If you suspect there is something embedded*") is separated in a clear way from the actions.
- The instructions for actions to be undertaken have a clear relative order, and it is obvious which is the first, which is the last, and which ones need to be taken in parallel.
- The part of the text to which the additional information ("*Explanation: In order to avoid putting pressure on the object itself.*") refers is clear.
- The more important parts of the text, i.e. the condition and the actions, are more visible than the additional information.
- There are no pronouns or omissions, and all objects are clear.

A more detailed description of the existing rules and a discussion of the specific high TC issues—which they address and how they are solved—will follow in the next two sections. Section 4.4.2 will provide a summary of the comparison between the LiSe and the CLCM guidelines.

4.3.2. The CLCM rules

As has been mentioned before, the CLCM guidelines contain over eighty rules. Although most of the examples in this thesis are concerned with applying them as a TS approach to simplify already-written texts, they can also be used as rules to follow for writing a new simple text from scratch (the application of TS to NLG has been described in Section 2.3.4.2.).

4.3.2.1. Types of rules per type of text complexity

The rules are of different types, according to the type of text difficulty or high text complexity that they address. In the scenario in which they are applied to produce a new text, they can be regarded as similar to the Natural Language Generation levels of generation. In fact, Controlled Text Simplification can be conceived of as NLG, in the sense that a text is generated according to the realisation of specific communicative goals. According to NLG, there are three general levels of planning leading to the realisation of a particular text: macroplanning, microplanning, and linguistic surface realisation levels (McDonald, 2000; Bateman and Zock, 2003; Vander Linden, 2000). In the same way, CLCM divides the rules into:

- Those which determine and structure the content and thus act on the macro-discourse structure of the text (information presentation, order and grouping).

- Those acting at the level of micro-discourse structure—not at the level of the whole text, but at the level of separate sections or paragraphs of it.

- Those acting at the sentence level and affecting concrete realisation choices at the word and sentence level.

The CLCM guidelines classification of rules divides them into five types: general, formatting, punctuation, syntactic, and lexical. A description of each of these types follows below.

The **General rules (G-rules)** define the purpose of the simplification and of the document, its structure, its contents, and the ways to order and group information, as well as imposing restrictions concerning the language of the whole document. These rules correspond to the first two levels of the NLG-classification (macro-discourse and micro-discourse). The G-rules are marked by the symbol "**G**", usually at the second or third position of the reference number (depending on whether the rule applies to the whole document or to a document sub-part). Examples of such rules are provided in Figure 4.11.

Rule restrictions	**Example**
Purpose of the simplification	**In_G_00:** Preserve the meaning and the information content of the original document.
Purpose of the document	**In_G_01:** This type of document is mainly aimed at the general audience (not specialists).
Document structure	**In_G_03:** The document should contain easily identifiable parts: (title of the document, sections, titles of sections and sub-sections, compulsory actions to be always done, conditions, instructions, lists, and comments).
How to group and order information	**In_G_09:** If there are distinguished situations: Identify the specific situations. Divide the blocks of instructions regarding the specific situations in subsections. Write first the most specific situation. Write the next more general situation. End with the most general situation. Write a title for each subsection, following the guidelines in Section "Guidelines for writing a title of a sub-section". **In_Cd_G_01**: If you have 2 or more alternative conditions: Start with the most specific one. End with the most general one.
The language of the whole document	**In_G_11:** Write in correct English. **In_G_12:** Begin sentences with a capital letter. **In_G_13:** Write only one piece of information per line. **In_G_14:** Write the cardinal numbers in figures.

Figure 4.11: Examples of G-rules.

As can be seen in Figure 4.11, the first column provides the type of restrictions of the rule, while the

second column provides one or more examples. Anyf incorrect and correct examples are omitted and only the rule reference number and its definition are provided, due to space limitations. There are thirty G-rules in total.

The **Formatting rules (F-rules)** define the graphical presentation of the document, including the formatting between and inside the different document sub-parts, specifying blank lines, font style, font size, indentation, etc., and are the first type of rules corresponding to the NLG concrete realisation level. The F-rules are marked by the symbol "**F**" and are in the same positions as the G and other types of rule markers. Examples of F-rules are provided in Figure 4.12.

Rule restrictions	**Example**
Formatting between document sub-parts	**In_F_T_02:** Jump 2 new lines after the title. **In_F_03:** Separate each block of instructions with a new line. **In_F_04:** Separate each group of conditions with a new line.
Internal formatting of document sub-parts	**In_Li_F_01:** Use the following formatting: Font style: regular, Font size: same as instructions and conditions Alignment: left<, Bullets: dashes (bullets), Indentation: +1. **In_I_F_01:** If the instruction is preceded by a condition: Indentation: +1. If not: Indentation: 0.

Figure 4.12: Examples of F-rules.

Like Figure 4.11, Figure 4.12 is composed of two columns, the first one indicating the type of restriction and the second one providing one or more examples. The incorrect and correct examples are omitted and only the rule reference number and its definition are provided, due to space limitations. There are only ten F-rules in total, as they are provided only as examples. More formatting rules will be added when there is a specific end-user and a concrete application.

The **Syntactic rules (S-rules)** impose restrictions on the syntactic realisation of the phrases and

sentences in the simplified documents. From the NLG perspective, they are the second type of rule dictating concrete realisation. The S-rules are marked by the symbol "**S**". The S-rules impose domain-independent and domain-dependent syntactic restrictions; a few examples are shown in Figure 4.13.

Rule restrictions	Example
Domain-independent restrictions	**In_S_01:** Use only the allowed syntactical structures. **In_S_02:** Avoid the forbidden syntactical structures on p.29.
Concrete, domain-dependent restrictions	**In_S_03:** Avoid demonstrative pronouns. **In_S_15:** If an adjective determines more than one noun: Repeat the adjective. **In_Li_S_01:** Use only NP with indefinite articles as elements of the list.

Figure 4.13: Examples of S-rules.

Figure 4.13 is also composed of two columns, the first one indicating the type of restriction and the second one providing one or more examples. The incorrect and correct examples are omitted and only the rule reference number and its definition are provided, due to space limitations. There are 24 S-rules in total. They are spread over the different document sub-parts, with further syntactic restrictions being provided at the end of the Guidelines (on pages 25 and 29).

The **Lexical rules (L-rules)** provide restrictions at the lexical level, and like the syntactic rules, can impose restrictions of a domain-independent and of a concrete, domain-dependent nature. They correspond to the NLG concrete realisation planning level. The L-rules are marked by the symbol "**L**". Examples of the L-rules are provided in Figure 4.14.

Rule restrictions	Example
Domain-independent restrictions	**In_L_01:** Choose the words in accordance with the lexical rules on p. 31. **In_L_04:** If possible: Use the alternative expressions in the dictionary on p.33.
Concrete, domain-dependent restrictions	**In_L_06:** If possible: Avoid acronyms and abbreviations. **In_I_L_01:** If possible: Use discourse connectives (e.g. first, second, next, then, finally).

Figure 4.14: Examples of L-rules.

Like the previous tables, Figure 4.14 is composed of two columns, the first one indicating the type of restriction and the second one providing one or more examples. The incorrect and correct examples are again omitted and only the rule reference number and its definition are provided, due to space limitations. There are seven L-rules in total, but further lexical restriction sections are provided at the end of the Guidelines (on page 31).

The final type of rules are the **Punctuation rules (P-rules)**, which impose restrictions on the use of punctuation marks in the document in domain-independent and domain-dependent cases. For this reason, they also correspond to the NLG concrete realisation level. The P-rules are marked by the symbol "**P**". Some examples of P-rules are provided in Figure 4.15.

Rule restrictions	Example
Domain-independent restrictions	**In_P_02:** Put the proper punctuation sign at the end of each line, as defined for every document part (instructions, conditions, etc.). **In_P_03:** Write a colon after the following elements: "If possible", "If not", "Perform the following actions simultaneously", "These are the instructions to follow", comments markers, conditions, instructions, followed by a list, elements of lists, followed by instructions.
Concrete, domain-dependent restrictions	**In_P_01:** If you make reference to a specific document: Put the document title in quotation marks. **In_T_P_01:** Avoid any punctuation signs at the end of the titles.

Figure 4.15: Examples of P-rules.

As can be seen in Figure 4.15, a punctuation rule can either refer to the whole document (domain-independent restrictions) or to specific elements (concrete, domain-dependent restrictions). There are thirteen P-rules in the guidelines.

4.3.2.2. Types of rules per CLCM guidelines section

As has been mentioned before, the CLCM guidelines map the existing *Instructions for the General Population* to a specific structure composed by a set of document sub-parts. Section 4.3.2. has already listed these sub-parts, which are: title of the document, titles of sections and sub-sections, instructions, conditions, comments, and lists of items. Rule *In_G_06* (page 5 of Appendix B) defines which of these elements are compulsory and must represent sub-parts of each document, and which are optional. A more detailed definition of these sub-parts follows below:

The **Title** is very important, as it specifies the topic of the whole document. It is important to separate it graphically from the other elements, and to make it short and meaningful. According to rule *In_G_06*, the title of the document is compulsory.

As the writer must identify the separate sub-situations and separate them into sub-sections, the **Titles of Sections or Sub-Sections** are also important, as they will help the user to easily identify the concrete situation for which she/he needs instructions. If there are specific situations large enough to be placed in separate sections, the titles of the sections or sub-sections (in case they are embedded) are compulsory.

The **Instructions** contain the main information, i.e. they list the actions to be undertaken during an emergency. For this reason they must be easily identifiable, short, and straightforward to understand. As the instructions represent the main information carried by this type of document, they are compulsory.

The **Conditions** are important because they specify which actions need to be undertaken under which circumstances. For this reason, they have to be placed before the Instructions. In addition, the instructions referring to a certain condition are indented (this has been seen in the rewritten example in Section 4.3.10). The conditions are optional, but their presence is preferable.

The **Lists** visually improve the understanding of enumerations. As it is known that more than seven elements are generally difficult to remember (*Miller's law*, Harley, 2008), it is advised to keep list items below this number. The lists are also optional, except when there are enumerations of more than two elements.

The **Comments** are the least important elements of the text. For this reason, they are optional. According to CLCM, there are two types of Comments–ones which can be put in the beginning of a document, such as a **Definition**, the **Target Audience**, or a **Reference to another document**; and those which can be put after a Condition or an Instruction, such as the **Aim**, an **Explanation**, an **Exception**, or an **Example**. Particular types of comments also include the Warnings, which can be written in order to warn about a dangerous situation, and which have a higher weight than the other types of comments. All of the comments are optional, except if there are important warnings.

As has been said before, the different rules characterize different document sub-parts. Table 4.3 provides the distribution of types of rules per document sub-parts.

Guidelines Section	**Types of rules**				
	G	**F**	**S**	**L**	**P**
General rules introducing the document	7	-	-	-	-
Rules for step-by-step document writing	9	2	17	6	3
Rules for the title	1	2	2	-	1
Rules for the titles of sections and sub-sections (same as for the title except the F-rules)	1	2	2	-	1
Rules for the conditions	4	-	1	-	2
Rules for the instructions	6	1	-	1	2
Rules for the lists	1	1	1	-	3
Rules for the comments	1	2	1	-	1

Table 4.3: Distribution of rule types per Guidelines sections.

The first column of Table 4.3 defines the CLCM guidelines section, while the other five columns contain the number of rules for each type (**G** meaning General rules, **F** meaning Formatting rules, **S** meaning Syntactic rules, **L** indicating Lexical rules, and **P** indicating Punctuation rules). The symbol "-" indicates that there are no rules of this type in this section. As can be seen, the rules introducing the purpose and the structure of the document are only of the G-type. The F-rules are,

as mentioned before, the least in number. The section explaining how to compose or rewrite a simple document contains the most rules. As the rules for the title of the document and the ones for the titles of the sections and sub-sections are the same (except of the formatting rules), they are of the same numbers. The rules for the Conditions have mostly G-rules (the majority of which define how to combine multiple conditions) and no F- or L-rules. In fact, almost all Lexical rules are in the step-by-step instructions for document writing, while most of the other sections, except the Instructions, do not have any of them. There are no syntactic rules for writing instructions and similarly to the case of Conditions, most of their G-rules define how to combine multiple parallel or alternative instructions. It can also be seen that the Lists and the Comments contain the lowest number of rules, although the number is sufficient enough for defining them.

4.3.2.3. Rules presentation

The rules contain compulsory and optional elements. The compulsory elements of each rule are its reference number (already described in Section 4.3.1.1) and the rule's definition. The optional elements of a rule are an incorrect and a correct example, and additional Comments. In order to make the Guidelines as comprehensible as possible, an attempt has been made to provide all rules with examples, and, if possible, with explanations. A screenshot of a complete rule is provided in Figure 4.16, while an example of a partially complete rule is shown in Figure 4.17.

Figure 4.16: Example of a complete rule.

Figure 4.17: Example of a partially complete rule.

As can be seen from Figure 4.16, the rule lists two examples–an incorrectly written one (on the left) and a correctly written one (on the right). Figure 4.17 shows that the partially complete example does not list any comment. This has been done in order to avoid providing overly technical comments for the writer, as the writer is considered to not be a linguist. In addition, the rules employ the allowed syntactic structures listed at the end of the guidelines, as has already been explained in Section 4.3.1. The list of rules for each specific section follows a particular order, allowing writing the document of the specific sub-part in a step-by-step fashion. For example, the section introducing the writing of the whole document starts with the following three rules:

- **In_G_07:** Write the title according to the guidelines in Section “Guidelines for writing a title”.
- **In_F_T_02:** Jump 2 new lines after the title.
- **In_G_08:** If there is a specific audience:
- Write “Target audience: target audience”.

Then, the explanations proceed with grouping the information:

- **In_G_09:** If there are distinguished situations: ...
- **In_G_10:** If there are sub-situations: ...

The rules defining the specific document sub-parts also attempt to follow this order of presentation. The next Section will discuss the high TC issues which are addressed by the CLCM rules.

4.3.3. High TC issues addressed by the CLCM rules

Table 4.4 lists the high text complexity issues addressed by the CLCM rules, comparing them with the high text complexity issues introduced in Section 2.1.3 and noting those whose frequency of occurrence was observed in the Crisis Management Corpus in Chapter 3. Due to the fact that CLCM addresses specifically the document type *Instructions for the General Population*, the attention is focussed on the rules which address the high TC issues affecting concretely this document type (a list provided in Section 3.3.2.1). These rules and issues are listed in **bold**.

Linguistic type	High Text Complexity issues presented in Chapter 2	High Text Complexity issues measured in Chapter 3	Rules
Lexical	**Rich vocabulary, percentage of different words**	**measured**	**In_L_01, In_L_02, In_L_03**
Lexical	**Long words, words with 2+ syllables**	**measured**	**In_L_01, In_L_02, In_L_03**
Lexical	Infrequent, technical terms	not measured	In_L_01, In_L_02, In_L_03, In_L_04
Lexical	**Ambiguous words**	**measured**	**In_G_14, In_G_15, In_L_02, In_L_03, In_L_05, In_L_06, In_S_09, In_Cm_G_01**
Lexical	**Vague quantifiers**	**measured**	**In_L_01, In_L_02, In_L_03 (not evaluated)**
Lexical	Words with high age-of-acquisition	not measured	In_L_03 (not evaluated)
Lexical	Abstract concepts	not measured	not addressed
Lexical	Words with high orthographic neighbourhood size	not measured	In_L_03 (not evaluated)
Lexical	**Inconsistent terminology, use of synonyms**	**measured**	**In_L_01, In_L_02, In_L_03**

Lexical	Figurative language	not measured	In_L_03 (not evaluated)
Syntactic	Long sentences, number of words in sentences	not measured	In_S_01, In_S_02, In_S_08, In_S_10, In_S_11, In_S_12, In_T_S_01, In_Cd_S_01, In_Cm_S_01
Syntactic	**Complicated syntax, convoluted structures**	**measured**	**In_S_01,** **In_S_02,** **In_S_08,** **In_S_10,** **In_S_11,** **In_S_12,** **In_S_13,** **In_S_14,** **In_S_15,** **In_S_16,** **In_S_17,** **In_T_S_01,** **In_Cd_S_01,** **In_Cm_G_01,** **In_Cm_S_01**
Syntactic	Too much information to remember	measured	In_G_13, In_Cd_G_01, In_Cd_G_02, In_Cd_G_03, In_Cd_G_04, In_Cd_P_02, In_I_G_03, In_I_G_04, In_I_G_05, In_I_G_06, In_I_P_02, In_Li_G_01, In_Li_P_03, In_Li_S_01
Syntactic	Passive voice	not measured	In_S_06
Syntactic	Negative constructions	not measured	In_S_07
Discourse	**Pronouns with unclear reference**	**measured**	**In_S_03,** **In_S_04,** **In_S_05**
Discourse	Illogical order	not measured	In_G_09, In_G_10, In_Cd_G_01, In_Cd_G_02, In_Cd_G_03, In_Cd_G_04, In_Cd_P_02, In_I_L_01 (not evaluated) In_I_G_02 (not evaluated)

Discourse	**Missing discourse connectives**	**measured**	**In_I_L_01**

Table 4.4: High text complexity issues addressed by the CLCM rules.

As can be seen from Table 4.4, the first column from left to right provides the linguistic type of each high TC issue, which can range from lexical to syntactic or discourse. The second column reflects the high TC issues presented in Chapter 2, Table 2.1, when the phenomenon of high text complexity was introduced, and the third column states which of these issues was measurable in the TC analysis of the Crisis Management Corpus in Chapter 3. Finally, the last column lists the rules which address each of these high TC issues. As can be seen, there are fewer high TC issues measured in Chapter 3 (column 3) than are presented in Chapter 2 (column 2). This is due to the fact (already explained in Chapters 2 and 3) that measuring some of these issues is either too results intensive or impossible, and that this set of measured high TC features already gives an estimation of the complexity of the corpus. As can be seen from the last column of the table, CLCM addresses more high TC issues than are measured in the TC analysis, which confirms that manual simplification can cover more high TC issues than automatic simplification, because both the automatic measurement of the presence of a certain linguistic phenomenon in text and its further automatic simplification imply in the first place the ability to detect it in the text. On the other hand, it can also be seen that some of the high TC issues introduced in Chapter 2 either have not been addressed by the CLCM rules, or were not evaluated in Chapters 5-7.

For example, CLCM does not provide a re-writing rule for abstract concepts, since it has been considered that such concepts would unlikely appear in CM texts. Also, some of the Chapter 2 high TC issues have not been directly addressed (by concrete rules), but are rather listed in the CLCM Section containing the forbidden lexical expressions and referred to by the lexical rule **In_L_03** (Avoid the forbidden lexical expressions). These high TC issues include:

- Vague quantifiers
- Words with high age-of-acquisition
- Words with high orthographic neighbourhood size
- Figurative language

This was because they have been added in a newer version of the CL and for this reason have not been included in the CLCM evaluation described in Chapters 5-7, as well as the following two rules:

- **In_I_L_01:** If possible: Use discourse connectives (e.g. first, second, next, then, finally).
- **In_I_G_02:** Use consecutive numbers for marking consecutive instructions.

In addition to the rules which address the high TC issues presented in this thesis, the rules ensure the preservation of the meaning and the information content of the text (Example: **In_G_00**); clarify it by specifying the target audience (Examples: **In_G_01, 02, 08**) and any reference documents (Example: **In_P_01**), by structuring it (Examples: **In_G_03, 04, 05, 06, 07**), by re-organizing and grouping its contents in a logical way (Examples: **In_G_09, 10**), and by making it more graphically readable (Examples: **In_F_T_02, In_F_03, 04, In_I_F_01**); and ensure that it is written in correct English and in a language appropriate for the domain (Examples: **In_G_11, 12, In_P_02, 03**). All of the high TC issues addressed have been provided with rewriting examples to clarify their simplification approach, which can be consulted in the CLCM Guidelines attached as Appendix B. Some examples of them have been already shown in the previous CLCM screenshots, for example in Figures 11 and 12, in which the verb and the preposition composing the phrasal verb have been

written together, while the passive voice has been transformed to active voice by replacing the object of the passive sentence with the subject of the active one. The next Section will introduce the result of adapting the MESSAGE CL technology and the transfer of the knowledge acquired while adapting CLCM to the Bulgarian language.

4.3.4. Adapting CLCM to Bulgarian

In order to test the MESSAGE project hypothesis that its technology is easily transferrable to other domains and languages, an experiment has been conducted in an attempt to transfer the CL technology from French and English to Bulgarian. The target documents were again instructions in the CM domain. The experiment took place during the training of the network of linguists (one of the deliverables of MESSAGE project – as stated in Section 4.2.1). The transfer to Bulgarian was done according to the specifications of the "Add MS kit"[35]. The work on the Bulgarian CL was conducted in collaboration with Ms. R. Margova[36]. The author's contribution for Bulgarian was the application of the Add MS kit to Bulgarian, while the second author's contribution was the concrete realisation in Bulgarian. The choice of the Bulgarian language was motivated by the following reasons:

- No CL has ever been created for Bulgarian, while many CLs have been developed for other languages (listed in Section 2.2).

- The research was conducted during a period of an increasing number of emergency situations with fatal consequences for the masses (Temnikova and Margova, 2009), due to

35 Available at http://message-project.univ-fcomte.fr/addmskit-en.html. Last accessed on January 30th 2011.
36 A linguist-specialist from the geodesic Journal "Geomedia", available at http://www.geomedia.bg/, last accessed on January 30th, 2011.

which a new Bulgarian Ministry of Emergency Situations has been created.

- The existing stock of emergency instructions was very limited, and the language in which they were written either followed an old style typical for the communist era (before the 1980s), or was translated from other languages (mainly from English), directly inheriting the source texts' characteristics without adapting them to the target language.

The guidelines for Bulgarian were developed on the basis of a collected corpus of emergency instructions. It was very restricted and consisted of both documents for specialists and for the general population. Due to the small size of the corpus, the CL for Bulgarian was also based on parallels with two other CLs for Greek and Polish, which were considered to be the closest to Bulgarian. The Bulgarian language can be considered similar to Greek and Polish because it is an Indo-European language, specifically a member of the Southern branch of the Slavic languages, and it has characteristics of both the Slavic and Balkan languages (of which Polish and Greek, respectively, are representatives). Due to the fact that it shares characteristics with both Slavic and Balkan languages, the Bulgarian language exhibits rich complexity at all levels, and its grammar, lexicon, and morphology can be considered much more complex than that of English (Боядиев, et al., 1999). For this reason, it constitutes a further linguistic challenge for building a TS approach. For example, in contrast with English, Bulgarian exhibits complexity and ambiguity at the morphological level. For this reason, to the original TC levels which CLCM contains rules to address (lexical, syntactic, and discourse), Bulgarian adds high TC issues at the morphological level. Due to the fact the Bulgarian CL is a prototype, discourse features have not been addressed. Examples of the high TC issues identified and the solutions designed by the Bulgarian linguist follow below:

High lexical TC issues

- Replace archaisms and clichés with an alternative synonym. For example, “служителите от пожарната” (the state employees of the fire brigade) → “пожарникари” (fire-fighters).
- Replace long words with synonyms composed of not more than one root and one suffix.
- Replace figurative language and metaphors.

High morphological TC issues

- Replace the short forms of the possessive pronouns with their full forms, in order to avoid ambiguity.
- Use the positive imperative with perfective verbs and the negative imperative with imperfective verbs, in order to avoid complexity and ambiguity.
- Replace present participles with a conjugated verb.
- Omit interjections.

High syntactic TC issues

- Split the condition from the instruction.
- Place the condition before the instruction using the expression “При:”.

Formatting rules

- Follow the formatting rules of CLCM in order to ensure uniformity to the writing standards

of the MESSAGE project.

The results of adapting CLCM to Bulgarian and applying the resulting rules to real-life instructions showed that the MESSAGE controlled language technology is highly domain- and language-specific, but the transfer is possible. The results also showed that it is less difficult and less time-consuming to adapt LiSe to English than to adapt LiSe and CLCM to Bulgarian. This can be explained by the fact that English and French are closer languages to each other than English and French are to Bulgarian. The created CL prototype has been presented to CM domain specialists [37] and received a few pragmatic corrections, but also a highly positive overall feedback regarding the information-grouping and linguistic decisions. An example of a resulting instruction rewritten using these rules is shown in Figure 4.18.

37 During a talk given at the Bulgarian Academy of Sciences in April, 2010.

Original	Re-written
В случаи на аварии и на извънредни ситуации, които могат да възникнат на място или по време на превоз, членовете и екипажът на превозното средство, са длъжни да предприемат следните мерки, от гледна точка на тяхната безопасност, и практическа възможност: - задействайте ръчната спирачка, изключете двигателя и откачете акумулаторните батерии,	При аварии и извънредни ситуации: -дръпнете ръчната спирачка, -изключете двигателя, -откачете акумулаторните батерии.

Figure 4.18: Example of a simplified instruction in Bulgarian.

Figure 4.18 shows that the first column contains the original instruction, while the second column contains the simplified or rewritten instruction following the CL rules. The translation of the original and rewritten examples is provided below:

Original: "*In cases of accidents and emergency situations, which may occur on site or during transportation, the members of the team and the crew members of the vehicle must follow the following measures, in relation to their safety and in terms of the practical possibility:*

- trigger the hand-operated brake, turn off the engine and disconnect the storage battery."

Simplified/Rewritten: "*In case of an accident or an emergency:*

- pull the hand-brake,

- switch off the engine,

- remove the battery.

It can be seen that after simplification, the text length is much smaller (45%, as reported in Temnikova and Margova, 2009) and the text is much clearer. The major problems in the original instructions were created by the introductory sentence, which provides the condition for the

instructions that follow. The rewritten example shows, in a way similar to the Plain English Campaign example provided in Section 2.1.1, that a long and convoluted sentence can be rewritten to a short one, while preserving its meaning and information content. More details about the linguistic motivations of the Bulgarian CL and the subsequent solutions can be found in Temnikova and Margova (2009). The next section will provide a comparison of CLCM with other CLs.

4.4. CLCM Characterisation

This section aims to provide an extensive characterisation of CLCM by comparing it to the controlled languages listed in Section 2.3.2, the CL from which it originated (LiSe, Renahy et al., 2010), and to describe it according to an existing CL specification with the aim of providing a means to compare different CLs (CNL 2009 specification, Section 4.4.3).

4.4.1. Comparison of CLCM with other controlled languages

As was introduced in Chapter 2, the existing CLs can be classified according to several criteria. These criteria include purpose, types of rules, and degree of automation. A few words reviewing these classifications, together with a statement about how CLCM is situated according to them, follows below.

In relation to their purpose, CLs can address human readers, machine applications, or a mixture of these two. In this way they can be defined as human-orientated, machine-orientated, and mixed-purpose CLs. Like LiSe, CLCM is a mixed-purpose CL, as although it is designed mainly to improve human comprehension of written text in emergency situations, it also aims to ensure good translation results, since this is very important in the modern global world. As such a CL, it is different from the formal-logic-based ones (described in Section 2.3.2 by having more free-text rules, but it is also different from the human-only CLs, described in Section 2.3.2.1, as it has more formal rules (constituted by a reference number, definition, and incorrect and correct examples).

In relation to the type of rules, a CL can have either prescriptive or proscriptive rules or construction/interpretation/paraphrasing rules. Some CLs have only a subset of these types of rules. As a reminder, prescriptive rules list the allowed expressions, proscriptive ones list the forbidden expressions, construction rules provide lexical and syntactic constructions, and interpretation rules assign unique interpretations of lexical and syntactic expressions, while paraphrasing rules provide paraphrases. As has been seen in Section 4.3.2, CLCM has all of these types of rules.

In relation to the degree of automation, a CL could be restricted only to manual resources, or could be provided with computer-aiding applications, or could be embedded in fully automatic systems.

With respect to this axis of classification, CLCM provides only manual resources, but a few steps towards determining the user requirements for a semi-automatic writing aid have been done in Chapter 7. An important characteristic of CLCM which distinguishes it from most CLs and TS approaches is that its clear aim is to preserve the meaning and the information content of the text to be simplified, while many approaches remove information that they consider redundant according to subjective criteria.

4.4.2. Comparison of CLCM with LiSe

The controlled language LiSe (Renahy et al., 2010) has been already described in Section 2.3.2.3 as an example of a mixed-purpose CL. Since CLCM has been developed from LiSe, a detailed comparison of them is needed in order to outline the novel aspects of CLCM. This section will describe the similarities and the differences between CLCM and LiSe.

4.4.2.1. Similarities between CLCM and LiSe

CLCM and LiSe share similarities in terms of the structure of the simplified documents, the contents and structure of their guidelines, and the set of rules. A list of similarities follows below:

- Both CLCM and LiSe map their target documents to a very similar document structure, and in this way the output simplified texts are composed by the same document sub-parts (except for a few differences listed below), and are characterized by the same formatting.

- The document sub-parts shared by CLCM and LiSe are the following: the title of the document, the titles of the sections or sub-sections, conditions, instructions, lists, some

comments (such as target readers; reference documents in the beginning of the document; and aim, explanation and exception following a condition or an instruction).

- CLCM was developed starting from the original structure of the LiSe guidelines; for this reason, their guidelines share the same elements, namely:
  - A table of contents listing the main document sections
  - A description of the rules' notation
  - A list of general rules for the whole document
  - Concrete rules for specific document sub-parts
  - Definitions of the allowed syntactic structures used in the various rules
  - An example of a rewritten text
- Most of the CLCM rules are the same as the LiSe rules, but have been adapted for English. A few new rules have been added after consultation with the psycholinguistic literature. The new rules are listed below. Due to the limited information published about LiSe, no examples can be cited. Some examples can be consulted in the "Extracts from Writing Manuals" document available online on the MESSAGE Project website. An example of such pairs of rules is provided in Figures 4.19 and 4.20.

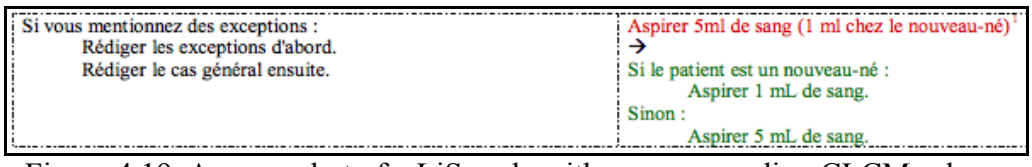


Figure 4.19: A screenshot of a LiSe rule with a corresponding CLCM rule.

The translation of the LiSe rule is as follows:

- Part on the left: "*If you mention any exceptions: Write first the exceptions. Then write the*

general situation after them."

- Part on the right: "*Draw 5ml of blood (1 ml with a newborn) → If the patient is a newborn: Draw 1ml of blood. If not: Draw 5ml of blood.*"

Figure 4.20 shows the rendering of this rule in CLCM, which has been translated, adapted to English, and expanded.

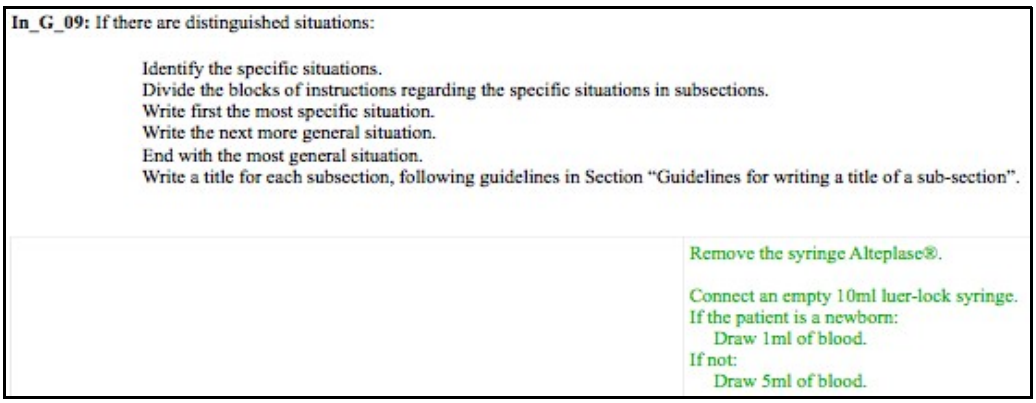


Figure 4.20: A screenshot of a CLCM rule with a corresponding LiSe rule.

In Figures 4.19 and 4.20 the same rule can be seen – the rule specifying that if there are specific situations, they must be listed first, before the more general ones, in order to avoid dangerous consequences. As can be seen, the example has been translated from French in order to ensure consistency between the MESSAGE Project controlled languages, but it can also be seen that the CLCM rule definition is much more elaborated than the LiSe one, at least in this version of LiSe.

- In order to ensure uniformity of the writing standards and facilitate transfer between the MESSAGE project CLs, the notation of the CLCM rules follows the notation of the LiSe rules; namely, each rule is characterized by a unique rule reference number, which is composed of a set of symbols indicating the type of document and the document sub-part to which the rule refers, as well as by a consecutive number.

4.4.2.2. Differences between CLCM and LiSe

The differences between CLCM and LiSe concern their target language, the target documents, the guidelines structure and additional resources, the set of rules, and the rules' visual presentation, as well as the document types participating in the rules' notation. These differences are discussed below.

- LiSe is developed for the French language, while CLCM is developed for English.

- LiSe addresses *Medical Protocols* and *Emergency Messages and Alerts* (Renahy, 2011), while CLCM addresses *Instructions for the General Population.*

- In terms of resources, in its unpublished and highly confidential version, LiSe features extensive sets (sometimes hundreds) of manually collected allowed syntactic and semantic structures (such as noun phrases and verb phrases), as well as adjectives and abbreviations together with their formal definitions. In contrast, CLCM, in order to enhance clarity for the writer of simple texts, enriches the structure of the LiSe guidelines with the following additional resources:

    - A grammatical term mini-dictionary
    - General rules valid for the whole document
    - A list of forbidden syntactic structures
    - A list of lexical rules
    - A list of forbidden lexical expressions
    - A domain dictionary with definitions for the lay reader

- A step-by-step rewriting example

- CLCM changes the rule notation by adding additional document type classifications. LiSe mentions only one type of document (*Protocols for specialists*, abbreviated as Pt). CLCM also has one main type of document (*Instructions for the General Population*, abbreviated In), but unlike LiSe, on the one hand it also defines the document types *Protocols* and *Alerts and Messages*, in order to make it possible to extend the CL to more types of CM documents, and on the other hand, CLCM distinguishes between different kinds of documents dedicated to addressing a different stage of an emergency situation ("during an emergency" being considered to have a higher importance, while "before and after an emergency" being considered to have a lower importance), as well as documents addressing different target readers (specialists versus the general population). According to these distinctions, CLCM distinguishes the following types of documents, characterized by the following notations, examples of which follow:

  - If the context does not require any distinction between the stages of the emergency situation and the target readers:

    Pr_ (protocol)

    In_ (instructions)

    Ame_ (alerts and messages)

  - If the context requires a distinction between the stages of the emergency situation but not between the target readers:

    PrNorm_ (protocol for before/after an emergency)

    PrDur_ (protocol for during an emergency)

InNorm_ (instructions for before/after an emergency)

InDur_ (instructions for during an emergency)

- If the context requires a distinction between the target readers but not the stages of the emergency situation:

AmeS_ (alerts and messages for specialists)

AmeG_ (alerts and messages for the general population)

- If the context requires a distinction between both the stages of the emergency situation and the target readers:

InNormS_ (instructions for before or after an emergency for specialists)

InDurS_ (instructions for during an emergency for specialists)

InNormG_ (instructions for before or after an emergency for the general population)

InDurG_ (instructions for during an emergency for the general population)

- CLCM changes the presentation and visualisation of the rules in order to make it clearer and more user-friendly. This difference has been already seen in the distinction between Figure 4.19 (a LiSe rule screenshot) and Figure 4.20 (a CLCM rule screenshot), or even better, in a comparison between Figure 4.19 and Figure 4.16 (a screenshot of a complete CLCM rule). As can be seen, the difference consists of the location of the main rule elements:

- The definition of the rule in LiSe is placed on the left hand side, while in CLCM it is placed at the top.

- LiSe lists both the incorrectly written and the correctly written examples on the right hand side, with the incorrect example first, an arrow symbolizing the transformation and

the correct example next, while CLCM clearly separates the incorrect and the correct examples by placing them on the sides of a one-row table.

- If a comment following a condition or an instruction is available, LiSe lists it below the rule definition, while CLCM places it at the end of the whole rule, under the table containing the examples. This is done because the comments generally bear the least important information of the rule.

- CLCM introduces some differences in the types of comments. Like LiSe, CLCM has two types of comments: those listed in the beginning of the document and those that follow instructions or conditions. In relation to these two types, CLCM drops some kinds of comments and adds some new ones.

Table 4.5 provides a comparison of the types of comments of LiSe and CLCM, along with examples. Note that the terms in the table are named as they are named in the CLs' guidelines.

LiSe		**CLCM**	
Explanatory notes in the beginning of the document		**Comment notes in the beginning of the document**	
Types	**Examples**	**Types**	**Examples**
Target readers	Public visé : médecins généralistes.	Target audience	Target audience: parents.
Author	Auteur: Equipe linguistique du Centre Tesnière.	*Removed*	
Date	Date: 12/04/2008.	*Removed*	
Reference document	Consulter la fiche technique N°3 «Entretien des voies veineuses centrale». *Ref: DSSI/PGPS/PSKT/01/M/17/06/99.*	Reference	*Ref.: www.mi5.gov.uk*
Explanatory notes following a condition or an instruction		**Comment notes following a condition or an instruction**	
Types	**Examples**	**Types**	**Examples**
Aim	Verser de l'eau sur la personne. But: Refroidir les brûlures.	Aim	Tap on pipes. *Aim: This will help rescuers to hear you.*

Explanation	Ne pas enlever les vêtements brûlés. Explication: Les vêtements brûlés collent à la peau.	Explanation	Plan an escape route to follow at night. *Explanation: Most fire deaths and injuries occur while people are sleeping.*
Exception	Ne pas utiliser les pronoms personnels. Exception: *vous*.	Exception	Avoid personal pronouns. *Exception: The personal pronoun "You".*
Not existing		Example	**In_P_03:** Write a colon after the following elements: /a list of situations follows/ /an example of a list of items follow/ *Example: Elements of a list followed by instructions.*
Not existing		Definition	/instructions follow/ *Definition: Under shock means not responsive.*
Not existing		Warning	Warning: Specific situations exist!

Table 4.5: Changes in the types of comments in CLCM compared to LiSe.

As can be seen from Table 4.5, the comment types at the beginning of the document are listed first, and then those which follow a condition or an instruction. The types of comments found in LiSe and CLCM are listed in parallel, with an indication of whether CLCM removes ("*removed*") or adds ("*Not existing*") a new type of comment provided in the respective place of the other CL. As can be seen, CLCM removes the author and date information, since they are considered irrelevant in instructions for the general population. In contrast, CLCM adds the optional comment types "**Definition**", "**Example**", and "**Warning**" to the list of comments following a condition or an instruction. The purpose of "**Definition**" is to provide a definition of an important technical term, in case the use of the term is required rather than a lay equivalent. The comment type "**Example**" provides indications regarding the given example in case the rule is complex and covers several situations, while the purpose of the "**Warning**" type is to draw the attention of the reader to a dangerous situation. Due to the importance of this comment type, "**Warning**" has a different format (the font is of a red colour) from the other types of comments, which are considered less important, and for this reason are written in italics, in grey, and with smaller sized fonts.

The last change refers to the set of rules. In addition to the LiSe rules being adapted from French to English and some of the LiSe rules being dropped completely (because they are applicable to French but not to English), CLCM adds a few new rules, including:

- The rule stipulating preservation of the meaning and information content of the original document (**In_G_00**)
- The rules specifying the target readers of the document (**In_G_01** and **In_G_02**)
- The rules defining the compulsory and optional document sub-parts and their relative order (**In_G_03, In_G_04, In_G_05** and **In_G_06**)
- The rule to keep verb and preposition together in phrasal verbs (**In_L_05**)
- The rule concerning how to treat acronyms and abbreviations (**In_L_06**)
- The rule stipulating using the Present Participle only as a verb (**In_S_09**)
- The rule suggesting the use of discourse connectives (**In_I_L_01**)
- The rule suggesting numbering the instructions in consecutive order (**In_I_G_02**)
- The rule explicitly forbidding the use of omissions (**In_S_13**)
- The lexical rules concerning avoiding figurative language, words with high age-of-acquisition, words with high orthographic neighbourhood size, and vague quantifiers listed at the end of the guidelines
- Some of the syntactic structures to avoid listed at the end of the guidelines (such as garden path sentences and ambiguous coordination)

The next section will describe CLCM by using an existing controlled language evaluation framework.

4.4.3. Description of CLCM according to the CNL 2009 specifications

As has already been shown in Chapter 2, the existing CLs are very different with respect to their purposes and their restrictions, and there are no existing common frameworks or evaluation methodologies allowing their comparison. For this reason, during the last Controlled Natural Language workshop, which took place in 2009 on Marettimo Island, Italy, it was decided to build a common framework for CLs, to make it possible to evaluate them within a common framework. Although this workshop and thus its evaluation framework concerns CLs having a formal logical basis, CLCM has been evaluated with it in an attempt to allow comparison with other CLs. The evaluation or CL description framework has been described in Wyner et al. (2009), who specifies five types of CNL properties. Four of these main groups of properties were taken into consideration in describing CLCM (generic properties, design properties, linguistic properties, and application properties), while the fifth one (relationships and evaluation) has not been taken into consideration, because it concerns a comparison between CLs. Table 4.6 lists the CLCM characteristics according to this specification. The information characterizing CLCM is enriched with information about the chapters and sections of this thesis where this has been mentioned.

1. Generic Properties		**Chapters and Sections**
Who are the intended users?	Non-specialists in emergency situations.	Section 4.1
What are the purposes?	To reduce complexity and to enhance comprehensibility in first place and possibly to improve human and machine translatability of emergency instructions.	Section 1.3, 4.5
Is the language domain dependent or independent?	The language is domain and language dependent.	Chapter 4
2. Design Properties		
Is the language easy to describe, teach and learn?	Feedback from users and trainings shows that it is.	Chapter 7
Is the language easy to read?	Yes.	Chapter 7
Is the language easy to write?	Yes, it is a naturally-sounding language.	Chapter 7
Is the language easy to understand?	Yes.	Chapter 5
Is the language predictable and unambiguous?	It tends to be.	Chapter 7
Is the language formally or informally defined?	Primarily informally, except for the syntactic constructions.	Section 4.3
How are semantic restrictions handled?	It is specified that each concept must be used with one pre-defined meaning.	Chapter 4
Are statements translated into logic?	No.	N/A
How is the CNL evaluated?	In terms of its impact on: High text complexity Text comprehensibility Manual translation Machine translation Users acceptability	Chapter 5, Chapter 6, Chapter 7
Is there a mapping to some graphical representation, e.g. conceptual graphs?	No.	N/A
Is the design of the language psycholinguistically motivated?	Yes.	Chapter 4
Is there an explicit statement of the syntactic and semantic theory which underwrites the language?	No.	N/A

Is the CNL easily and systematically extensible (adding lexical, morphological, syntactic, and semantic elements or components)?	Yes.	Section 4.3.4
3. Linguistic Properties		
What corpus (if any) is one using to judge which linguistic forms to include in the language?	Crisis Management corpus.	Chapter 3
What linguistic literature or theory (if any) is one using to justify the linguistic properties of the language?	Harley (2008)	Chapter 2
What classes of nouns, verbs, adjectives, adverbs, quantifiers, etc. are supported?	All English language classes, except pronouns and vague quantifiers.	Chapter 4
Does the lexicon support polysemy or only monosemy? How is polysemy resolved relative to context?	Monosemy is strongly suggested whenever possible. Every concept has to be used with its pre-defined meaning.	Chapter 4
Is the language mono-lingual, or does it support multilinguality?	The language is monolingual but multilinguality is supported in terms of its improvement of human and machine translation.	Chapter 4, Chapter 6
What morphological word formation rules are supported?	All as in the English language.	Chapter 4
Are interrogative and imperative forms supported? Are they generated from assertions or must they be explicitly written?	Yes. They must be explicitly written.	Chapter 4
Are idioms and metaphors allowed?	No.	Chapter 4
Diathesis alternations (passive-actives, middles, ditransitives, causatives, inchoatives which signal the beginning of an action, and others). What inferential patterns are supported?	The passive voice is forbidden.	Chapter 4

Where we have synonymous syntactic forms (outside the scope of diathesis), which should we adopt? How should relationships between them be defined?	In the list of forbidden and allowed syntactic structures.	Chapter 4
Is there syntactic sugar, i.e. redundant expressions that make some expressions easier to state?	No, except in Bulgarian.	Section 4.3.3
Can there be discontinuous constituent structures, interruptions, or higher-level speech acts?	No, only standard word order is supported.	N/A
What sorts of query, relative clause, and sentence subordination markers are supported?	Relative clauses, except the restrictive relative clauses must be avoided.	Chapter 4
What sorts of subordinate clauses are supported?	Subordinated clauses must be avoided whenever possible.	Chapter 4
Is discourse supported?	Yes, in terms of information grouping and ordering, the use of discourse connectives and forbidding the use of pronouns.	Chapter 4
What aspects of anaphora are considered: times, locations, facts, propositions, and definite descriptions?	Pronominal anaphora must be avoided.	Chapter 4
4. Application Properties		
Are there automatic consistency checks?	No.	N/A
Are there automatic redundancy checks?	No.	N/A
Is there guidance on style?	Yes.	Chapter 4
What support tools are provided by the CNL?	Nothing automatic—only printed resources, such as: • A grammatical terms mini-dictionary • A list of allowed syntactic structures • A list of forbidden syntactic constructions • A list of lexical rules • A list of forbidden lexical expressions • A domain dictionary • A step-by-step rewriting example • A rewritten “before and after” example	Chapter 4
How is the language maintained and developed? Is the CNL proprietary or open?	The CNL is proprietary.	Chapter 1

Table 4.6: Characterisation of CLCM according to the CNL 2009 specifications.

As can be seen from Table 4.6, the four CNL properties are listed one after another, with the first column listing the questions relevant to each type of property and the second column providing the answers to these questions. As can be seen, the generic properties define the purposes of the CL. The design properties encompass the evaluation questions, or potential mapping to a formal language. The linguistic properties describe the CL from several linguistic perspectives, while the application properties ask questions about automatic applications supporting the current CL. Since in our case there are no automatic applications, this section has been used to describe the printed resources available for CLCM. It should also be noted that in comparison with the original set of properties, not all properties have been listed in Table 4.6. This is due to the fact that the omitted properties strictly define only formal logic-based CLs and are not applicable to CLCM. This comparison has been done with the hope of making CLCM easily comparable to other CLs. The next section will provide a comparison between CLCM as it exists for English and the prototype for Bulgarian developed during the MESSAGE Project.

4.5. Conclusions

This chapter has presented the proposed Text Simplification approach–the Controlled Language for Crisis Management. The chapter has provided a description of the project context, the documents for which it is designed, as well as an extensive presentation of the proposed CL for English, including its guidelines and an analysis of the rules. Finally, in order to provide a better view of the proposed TS approach, a comparison of it from three different perspectives has been provided. CLCM has been compared first to the CL from which it originated–LiSe, the CL for French. It has been shown that in addition to adapting it to English, CLCM enriched it with additional resources,

new rules, and better guideline visualisation. Next, CLCM has been described from the perspective of an existing CL specification framework, with the view to making it comparable to other CLs. The results have shown that CLCM fits the existing framework description and provides positive answers to most of the framework's questions. Finally, a transfer of CLCM to Bulgarian has been presented. The results have shown that the MESSAGE CL technology is highly language-specific, but that the document features can easily be adapted to new languages.

The design of CLCM has been based on a stable CL tradition (LiSe) in narrow collaboration with end-users, and its conformity to psycholinguistic findings on high TC issues hindering comprehension and enhancing human comprehension in emergency situations has been ensured. However, it is still important to investigate whether the simplified text exhibits reduction in text complexity and improvement in reading comprehension, as well as an improvement in additional tasks which are important for the CM domain, such as manual translation and machine translation, as well as examining whether simplifying according the CLCM guidelines poses difficulties to writers of simplified texts.

The next chapters will present an extensive evaluation of CLCM from several perspectives. As the second main aim of this thesis is to propose a method for enhancing emergency comprehension, Chapter 5 will investigate the hypothesis that text simplified according to the CLCM rules exhibits improvement in its primary features, such as text complexity and reading comprehension. Chapter 6 will test the hypothesis that the simplified text has a positive impact on additional, but also important for the domain extrinsic tasks such as manual and machine translation. Finally, Chapter 7 will conduct an in-depth evaluation of the manual text simplification process and will examine the difficulties that human writers encounter while simplifying text.

Chapter 5 – The Effect of CLCM Simplification on Reading Comprehension

Neither comprehension nor learning can take place in an atmosphere of anxiety.
(Rose Kennedy)

The aim of this chapter is to present the evaluation of CLCM from the first point of view – its impact on reading comprehension under stress – via what will be referred to as the "Online Reading Comprehension Experiment". Section 5.1 will provide the Introduction to the chapter and the motivation for the experiment. Section 5.2 will present the related work on evaluating Controlled Languages and the evaluation perspective taken by this thesis. Section 5.3 will describe the setting of the experiment, Section 5.4 will provide the evaluation results, and Section 5.5 will discuss the findings and present some critiques of the experiment, as well as ideas for related future work. Finally, Section 5.6 will present the conclusions.

5.1. Introduction

After the mixed-purpose (see Section 2.3.2.3) Controlled Language for Crisis Management (described in Chapter 4) was developed in order to address the simplification needs of the Crisis Management Corpus document type *Instructions for the General Population* (IGP) outlined in Chapter 3, the next step was to evaluate it. Since, as stated in Section 1.3, the main purpose of CLCM is to enhance comprehension, but, as stated in Section 4.4.3, it is also desirable that it has a positive impact on other important for the domain tasks, the evaluation focused on determining if CLCM has a positive impact on human reading comprehension under stress, and in addition, whether it has a positive impact on specific tasks important for the CM domain (such as translation). A final evaluation point of view is whether the way the CLCM was designed makes its application easy and what causes difficulties in using it to simplify texts. The aim of the present chapter and the two following ones is thus to present the evaluation of CLCM from multiple perspectives. While Chapter 6 will treat the evaluation of the impact of the CLCM simplification on manual and machine translation and Chapter 7 will investigate the internal process of manual text simplification, the present chapter will address the first and most important evaluation perspective of CLCM—its impact on reading comprehension. In order for this to be achieved, a large scale experiment, the *"Online Reading Comprehension Experiment"*, involving over one hundred human participants, and attempting to minimize any irrelevant variables, will be described. The experiment consisted of asking human volunteers to read four texts containing emergency instructions, two of which were original (complex) and two of which were simplified, and reply to a set of questions after each of the texts. In order to simulate (at least to a certain extent) an emergency situation and generate stress, the time for reading the texts was limited. The results are evaluated using two evaluation metrics, namely percentage of correct answers and time to provide correct answers.

Next, Section 5.2 will present the existing approaches in evaluating controlled languages.

5.2. Evaluating Controlled Languages

This section will introduce the related work on evaluating Controlled Languages (CL) as a language resource (Section 5.2.1), the view taken by this thesis, and the CLCM evaluation approach (Section 5.2.2).

5.2.1. Related work in controlled language evaluation

The related work in evaluating controlled languages can be classified into four different types of approach:

- Quality estimation of the controlled language re-writing: feedback from end-users, comparison of the output with other controlled languages
- Evaluation of the ease of writing in a controlled language
- Evaluation of the impact of the controlled language on human comprehension
- Evaluation of the impact of the controlled language on other tasks

controlled language for their needs (Renahy et al., 2010). Comparing the CL output with the outputs of other CL generated from the same input sentence (Pool, 2006) allows evaluation of whether all of the ambiguities or complexities in the input sentence were resolved. The shortcoming of the first kind of approach is that the free-text form is hardly objectively and numerically quantifiable, while the limitation of the second kind is that it does not give an evaluation of the absolute quality of the controlled language in question.

The evaluation of the ease of writing in a controlled language (Kuhn, 2010), to the knowledge of the author, has only been applied in formal controlled languages (i.e. those allowing mapping to formal languages or formal knowledge representations) and consists of presenting a task of formulating a statement in the CL with the use of a special tool. The shortcomings of this approach come from the fact that the limitations of the tool can affect the evaluation of the CL itself (Kuhn, 2010).

The approaches that involve evaluating the impact of the CL on human comprehension comprise paraphrases and ontographs (Kuhn, 2010), and, to the author's knowledge, are restricted only to formal controlled languages. The paraphrase approach (Kuhn, 2010) consists of a statement formulated in the controlled language and four paraphrases of it in a natural language. The participant is asked to chose only one of them whose meaning most corresponds to the meaning of the CL statement. The limitation of this approach is that it is not sure whether the participant would correctly understand the paraphrases, as they generally use ambiguous quantifiers and referents (Kuhn, 2010). Ontographs (Kuhn, 2010) are a novel approach designed for testing formal languages. They are schematic diagrams representing situations and the participants in the situations, as well as their actions and roles. The ontographs kit can be accessed at http://attempto.ifi.uzh.ch/site/docs/ontograph/#Kit[38]. The limitation of this approach is that the

38 Last accessed on March 6th, 2012.

diagrams are manually generated and require a large amount of time and effort for their production.

The impact of controlled languages on other tasks has been used as an indirect evaluation of the controlled language and has been applied for machine translation (Vassiliou et al, 2003; O'Brien & Roturier, 2007; Aikawa, et al, 2007) and database search (Cleverdon, 1977). Cleverdon (1977) evaluated the performance of a CL on the task of retrieving texts from a database by formulating the queries in natural language and in the CL. The results were in favour of the natural language, probably due to fact that the texts to be retrieved and there titles were written in natural language. The approaches that use MT and test the impact of the CL on it have applied several evaluation techniques, namely:

Manual rating of MT translations focussing on the translations of concrete linguistic issues in the input text (Vassiliou et al, 2003). The limitations of this approach are that it is very subjective and has a limited focus.

Post-editing evaluation using edit distance (O'Brien & Roturier, 2007; Aikawa et al, 2007), time (O'Brien & Roturier, 2007), CNA (O'Brien & Roturier, 2007), BLEU (Aikawa et al, 2007), and human rating of MT translation comprehensibility (O'Brien & Roturier, 2007) and acceptability (Aikawa et al, 2007). The limitations of these techniques will be discussed in Chapter 6.

5.2.2. Thesis evaluation perspective

On the basis of the existing work in CL evaluation, the approach taken to evaluate CLCM in this thesis is an innovative approach that aims to simulate the basic principle of evaluation of output systems, as defined in Hirshman and Mani (2001), i.e. by evaluating the CLCM resource on specific

tasks, i.e. by extrinsic measures.

The extrinsic measures applied consist of assessing the performance of the texts, simplified according to the CLCM guidelines, in measuring text complexity, reading comprehension and in the manual and machine translation tasks. In addition, evaluation of the cost involved in manually simplifying according to the CLCM guidelines is performed, by measuring the simplification speed and identifying the difficulties encountered while manually simplifying texts. The latter results can be used in future work as guidance regarding the implementation of an automatic system to facilitate manual simplification.

The evaluations of CLCM from the above described perspective are divided into three chapters and described respectively in Chapter 5 and Chapter 6 (extrinsic evaluation), and in Chapter 7 (cost evaluation). The original contributions of this thesis regarding CL evaluation are thus in the general view of evaluation, the method adopted for evaluating the CL impact on human comprehension (described further in this chapter), the novel MT post-editing cognitive evaluation approach (described in Chapter 6), and the method for evaluating the internal process of manual text simplification by providing objective numerical results.

Next, Section 5.3 will present the setting of the first evaluation experiment, namely the extrinsic evaluation of CLCM on reading comprehension.

5.3. Setting of the Experiment

In order to evaluate the impact of the CLCM simplification on reading comprehension, the *"Online Reading Comprehension Experiment"* was designed and conducted. The experiment consisted of

asking a large number of volunteers to read several texts containing emergency instructions in a limited amount of time and reply to a set of questions following them. Some of the texts were left as they were originally. They will be referred to as “complex” form now on. Some were simplified, according to the CLCM simplification rules. From now on, these texts will be referred to as “simplified.” The experiment was conducted online. A special interface was developed.

This section will describe the setting of the experiment. Section 5.3.1 will present the research hypotheses investigated, Section 5.3.2 will describe the running of the experiment, Section 5.3.3 will provide technical details regarding the design of the experiment, and Section 5.3.4 will discuss the experiment preparation and pilot experiments.

This experiment was conducted with the contributions of Dr. Constantin Orasan and Dr. Le An Ha. Their contributions consisted of implementing the web interface used in the experiment. Its functionalities will be described in Section 5.3.2.

5.3.1. Research hypotheses investigated

As stated above, the aim of the experiment was to evaluate the impact of CLCM on reading comprehension. Specifically, the goal of the experiment was to test the following research hypothesis:

The CLCM simplification has a positive impact on reading comprehension.

This hypothesis was tested by measuring the time employed by participants to correctly answer a set of questions following the emergency instruction texts and by calculating the percentage of

correct answers from all answers given to the questions of the particular text. The results for the complex and the simplified texts will be compared. The reason that it was decided to take into account only the correct answers and to ignore the incorrect and "*I don't know*" answers was that the aim of the CLCM simplification is to enhance comprehension of emergency instructions, and thus the ability of readers to identify and give the correct answers is a good measure of correct comprehension of the texts.

It is assumed that if CLCM has a positive impact on reading comprehension, then:

1. The percentage of correct answers given for the simplified text will be higher than the percentage of correct answers given for the complex text.
2. The time to recognize the correct answer and reply correctly to the questions about the simplified text will be significantly less than the time to recognize the correct answer and reply correctly to the questions about the complex text.

Next, Section 5.3.2 will describe the way in which the experiment was conducted.

5.3.2. Unrolling of the experiment

The experiment employed a specially developed web interface, the welcoming screen of which can be seen in Figure 5.1.

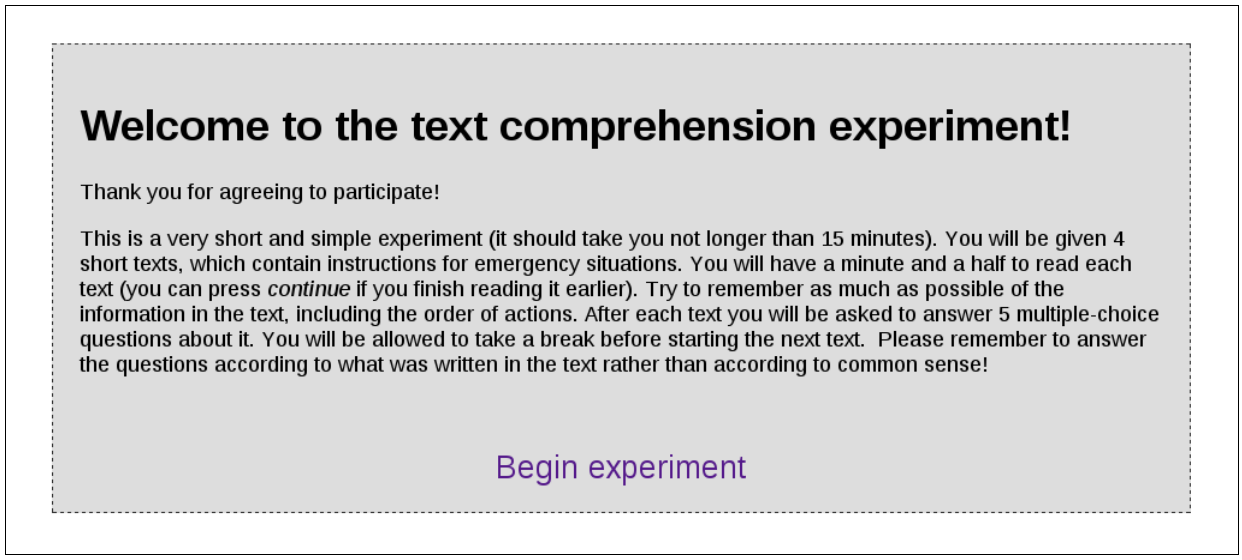


Figure 5.1: Welcoming screen of the web interface.

As can be seen in Figure 5.1, the web interface was very clear and user-friendly and the welcoming screen briefly informed the participant about the aim of the experiment, the way that the experiment would take place, the tasks of the participant, and the duration. The duration of the experiment was kept to be very short in order to ensure a large number of volunteers who would agree to participate in it. It was tested through a pilot experiment, described in Section 5.3.4. The description of the experiment provided in the welcoming screen followed the instructions for participants distributed via e-mail during the recruitment of volunteers. The instructions for participants are provided in Appendix C.

The experiment commenced after clicking on “Begin experiment”. First, the participant was prompted to enter detailed personal data, which would allow a subsequent data analysis based on different participants’ variables. The participant was reassured that his/her personal data would not be published or disclosed to third parties. The following data was collected:

- Name

- Age
- Sex (female/male)
- Occupation
- Background in Crisis Management (yes/no)
- Native language
- E-mail address
- Level of English – native/advanced/intermediate/beginner

After collecting demographic data, the test commenced. The experiment involved presenting four texts—two complex ones and two simplified ones—containing instructions for emergency situations. The simplified ones were manually simplified according to the CLCM guidelines, and questions about their contents were generated manually.

After the welcoming screen and before displaying the first text, an introductory message was displayed. The message was different for each text. The aim of these messages was to prepare the participant for the topic to come and to provide more concrete instructions, for example, to explain that the text would be displayed for a limited time and that the participant should read as quickly as possible and try to understand as much as possible, as well as to ignore his/her general knowledge of the topic. The four introductory messages are provided in Appendix C. At the end of each introductory message a link labelled “Continue” was provided, which the participant was invited to click when ready for the test.

After clicking on the link, the time-limited text was displayed. After the one and a half minute

period finished, the text disappeared and the first question tailored to this specific text was displayed. Unlike the text presentation, the task of answering the questions was not time-limited, but time to response was recorded.

Four or five questions were displayed after each text. They were in the form of multiple-choice questions. Each question was composed of a question and four answers. An example of a question is provided below:

Question 36 (Set 2):

According to the text, you should return to the house:

- *If you are told it is safe to do so.*
- *After calling the fire department.*
- *To avoid fire, electrocution or explosions.*
- *If you smell gas.*

The time for answering questions was recorded. Therefore, participants were instructed not to get distracted while replying to questions, and were promised that after each set of text plus questions, a break would follow. After replying to all questions following a specific text, the introductory message of the next text appeared, giving the participant the option of taking a break and continuing when ready to the next set of text plus questions. After finishing with all four texts plus questions, a "Thank you" screen was displayed. The welcome and goodbye texts are provided in Appendix C.

5.3.3. Technical setting of the experiment

Behind the clear and user-friendly interface, there was a complex automated set of functionalities, the aim of which was to record participants' personal data, answers to the questions, and time employed to provide answers, as well as selecting the texts and questions to appear. This section will describe the technical setting of the experiment. Section 5.3.3.1 will describe the texts used in the experiment, Section 5.3.3.2 will describe the development of the questions, Section 5.3.3.3 will present the method by which the texts were selected for viewing, Section 5.3.3.4 will describe the method of generating the order of the questions and answers, and finally, Section 5.3.3.5 will describe the method of recording participants' data. The latter was used for obtaining the results in Section 5.4. The ideas for the methods for generating the texts, questions, and answers, as well as the limited-time setting (described respectively in Sections 5.3.3.3, 5.3.3.4 and 5.3.3.5), were developed by the author of the present thesis, while the implementation was executed by Dr. Constantin Orasan and Dr. Le An Ha, together with the method of recording the participants' data and its implementation.

5.3.3.1. Text selection

The texts used in the online experiment were taken from the texts resulting from the manual simplification of Subject 1 (considered to be the most expert in simplifying) during the Text Simplification Task experiment, described later in Chapter 7. The texts from the experiment in Chapter 7 were shortened so as to be the same length, around 150-160 words, in order to not tire the participants and to ensure comparability. Table 5.1 provides the topics and length of the eight texts per pair of complex-simplified text, while their texts are provided in Appendix C.

Topic	Text	Size
How to clean your home and stop mold after a flood.	Complex Text 1	Words: 165 Characters: 903
	Simplified Text 1	Words: 174 Characters: 995
Precautions when returning home after a flood.	Complex Text 2	Words: 165 Characters: 917
	Simplified Text 2	Words: 146 Characters: 915
How to do personal cleaning and to dispose of contaminated clothing.	Complex Text 3	Words: 166 Characters: 963
	Simplified Text 3	Words: 153 Characters: 895
Protecting yourself after a volcanic eruption.	Complex Text 4	Words: 165 Characters: 1012
	Simplified Text 4	Words: 172 Characters: 1108

Table 5.1: Topics and sizes of the texts used in the experiment.

As can be seen in Table 5.1, the first column displays the text's topic, the second column displays the name of the text, and the third column displays its length in words and characters. As can be seen from Column 1, the topics are the same for each pair of complex-simplified texts, as the simplified texts originated from a manual simplification of the complex ones, which preserved the text content. As can be seen from Column 3, the lengths of the complex texts were made to be more or less the same, in order to ensure comparability. It was impossible to limit the lengths of the simplified texts, as they depended on the simplification output of the specific human simplifier.

5.3.3.2. Developing the questions

The questions used in the experiment were manually created on the basis of the four complex texts, described in Section 5.3.3.1. The four complex texts were analysed for presence of the following kinds of information:

- Lists of items

    - For example: "*Wear rubber boots, rubber gloves, and goggles when cleaning with bleach.*" (Sentence taken from Complex Text 1.)
- Order of actions
    - For example: "*To remove mold, mix 1 cup of bleach in 1 gallon of water, wash the item with the bleach mixture, scrub rough surfaces with a stiff brush, rinse the item with clean water, then dry it or leave it to dry.*" (Sentence taken from Complex Text 1.)
- Key details
    - For example: "*Never mix bleach and ammonia.*" (Sentence taken from Complex Text 1.)
- Conditions
    - For example: "*If your eyes are burning or your vision is blurred, rinse your eyes with plain water for 10 to 15 minutes.*" (Sentence taken from Complex Text 3.)
- Explanations
    - For example: "*Never mix bleach and ammonia. The fumes from the mixture could kill you.*" (Sentence taken from Complex Text 1.)

After the complex texts were analysed for the presence of these kinds of important information, it was determined whether the same information was preserved in the simplified texts and whether modifications to the original wording were made. If the same important information was preserved in the simplified texts, a question about it was created. In this way, the same set of questions could be asked about the complex text and the corresponding simplified text, which ensured comparability.

The questions, as already seen in Section 5.3.3.1, were in the form of multiple-choice questions, composed of a stem (the question itself), the correct answer, three incorrect answers (called

"distractors") and the answer "*I don't know.*" The multiple-choice form was selected as it is the most objective method of measuring comprehension (Gronlund, 1982), which only involves determining the amount of correct answers.

The questions were formulated according to some of the rules described in Gronlund (1982), namely:

- Design each item to measure an important learning outcome.
- Present a single, clearly formulated problem in the stem of the item.
- State the stem of the item in positive form, wherever possible.
- Make sure that there is only one correct answer.
- Make all distractors grammatically consistent with the correct answer, in order to avoid hints towards the correct answer.
- Avoid verbal cues that might enable participants to recognize the correct answer.
- Make all answers the same length, in order to avoid easy recognition of the correct answer.
- Make the distractors plausible and attractive to the uninformed.
- Avoid providing information in the stem or in one of the questions which may point to the correct answer of the same question or of another question.
- Use an efficient item format; make the stem and the answers clearly visible.

The next section will describe the methods followed in order to randomize the texts.

5.3.3.3. Text randomisation

Although there were eight texts, only four of them, two complex and two simplified, were displayed to an individual participant. In order to avoid an influence of the order of the texts on the results, three approaches were followed:

1. The texts were displayed in one of two alternating orders:
    - Simplified text – complex text – simplified text – complex text
    - Complex text – simplified text – complex text – simplified text
2. For each of the positions in each alternating order, the texts displayed were selected randomly. For example, participant 1 could get "Complex Text 1 – Simplified Text 2 – Complex Text 4 – Simplified Text 3", participant 2 might get "Complex Text 4 – Simplified Text 3 – Complex Text 1 – Simplified Text 2", and participant 3 might get "Simplified Text 1 – Complex Text 3 – Simplified Text 4 – Complex Text 2".

3. In addition, each text, whether complex or simplified, was shown to each participant only once, and after a complex text was shown, its simplified version was not shown, and vice-versa.

This method ensured that each participant was presented with a different combination of the texts, and in a different order. In order for this to be implemented, the eight texts were given unique numbers and were grouped into four sets, each containing a complex-simplified pair.

- Set 1: Text 1 (Complex) and Text 2 (Simplified)

- Set 2: Text 3 (Complex) and Text 4 (Simplified)
- Set 3: Text 5 (Complex) and Text 6 (Simplified)
- Set 4: Text 7 (Complex) and Text 8 (Simplified)

5.3.3.4. Display time for texts

Each text was displayed for a limited amount of time—one and a half minutes. The duration of the display of the texts was motivated by two reasons:

- The knowledge that the average reading speed of an adult is around 200-300 words per minute for reading with learning and understanding (Carver, 1992) and that the length of each of the texts is around 150-160 words.
- The findings based on results from the pilot experiment, which will be described in Section 5.3.4; it showed that experimental subjects employed on average one minute and a half to read the complex versions of the texts.

5.3.3.5. Question and answer randomisation

Similarly to the texts, in order to avoid order effects on comprehension, the questions were displayed in a random order. In order to ensure clear attribution of a question to the correct couple of complex-simplified texts, each question was given a unique number. Which question numbers correspond to which set of texts was then recorded. For example, the questions attributed to Set 1 were numbers 25, 27, 28, 29, and 33, while those attributed to Set 4 were 45, 48, 49, and 50.

Additionally, as per Gronlund (1982), it is desirable to place the correct answer in different positions, in order to not allow the participant to guess the pattern of the answers. For this reason, the position of the correct answer to each question was also randomized.

5.3.3.6. Recording experimental data

In order to effectively calculate the number of correct answers in the face of randomisation of answers, each answer received a reference number, similarly to the questions. In each question, the correct answer was associated with the value “0”, the distractors were associated with the numbers “1”, “2”, and “3”, and, finally, the answer “I don't know” was associated with the value “100”. In this way, independently of its position, each time the correct answer was selected the program recorded the number “0”, allowing in this way easy counting of the correct and incorrect answers.

As also mentioned earlier, the time taken to read the question, read the answers, and select an answer was also measured, for all given answers, correct or not. Further on, their associated values, described in the previous paragraph, helped with recognizing which time was recorded for which answer. The time employed to give an answer was measured in milliseconds.

The data was recorded in the following way: at the time that a participant, after entering his data and reading the text, read and provided an answer to a question, a line in the database was generated, including all of the information regarding this entry, this participant, this text, its set, the question, the given answer, the time employed, and whether the participant has completed the whole test (all four texts and their questions). An example of a recorded line, is provided in Table 5.2.

id	age	sex	occupat.	backgr.	lang.	name	e-mail	level	text	compl.	user	quest.	answer	time	set

1	24	f	Student	n	English	–	–	native	4	1	1	30	0	18695	1

Table 5.2: Information recorded per answer.

As can be seen in Table 5.2, the data recorded per answer was quite complex:

- the first column, "id", contained the number of the entry;
- the second column – the age of participant;
- the third column – his/her gender;
- the fourth column – his/her profession;
- the fifth column – whether he/she had experience in the CM domain;
- the sixth column showed his/her native language;
- the seventh and eighth columns contained the participant's name and e-mail address, respectively (omitted here for confidentiality reasons);
- the ninth column contained the level of English;
- the tenth column contained the number of the text (in this case "*4*", i.e. the simplified text from Set 2);
- the eleventh column indicated whether the whole test was complete ("*1*" meaning yes, "*0*" meaning no);
- the twelfth column contained the number of the user (in this case the first one);
- the thirteenth recorded the answer (in this case "*0*", the correct one);

- the next column displayed the time in milliseconds;
- the last column contained the number of the set.

For programming reasons the sets 1-4 were assigned the numbers from 0 to 3.

The test made by each participant was thus recorded in one to nineteen rows (depending on how much of the test was done). In each of these rows the positions from two to nine and eleven contained the same information, as they corresponded to the description of the user.

5.3.4. Preparation of the experiment: pilot experiments and advertisement

Before being launched, the experiment passed through two testing stages in the form of pilot experiments. The first pilot experiment took place after choosing the texts, analysing them for important segments of information, composing the questions and answers, and consulting the psycholinguist. Its aim was to test the quality of the questions and the selected answers and to receive feedback about the experiment as a whole. The experiment consisted of asking five volunteers to read the complex versions of the four texts in a printed form, to reply to the prepared questions, and to provide feedback about the flaws of the experiment. The participants were all male, ranging between third-year undergraduate students and post-doctoral fellows in computer science and NLP. Their level of English ranged from beginner to advanced. The instructions given to the participants were:

Task:

- *Measure the time taken for reading each text.*
- *Measure the time taken for answering all of the questions for one text.*
- *Read each text.*
- *Mark the correct answer after each question.*
- *Think about any remarks about the quality of the texts/questions.*

The results from this pilot experiment made it possible to check how much time it took to read the texts, in order to estimate how much time to restrict the reading of the texts to in the actual experiment. It also made it possible to estimate how much time the whole experiment would have taken. This pilot experiment also made it possible to identify any issues in the questions, such as questions which were too hard, questions which were too easy, and overlapping answers or hints in the questions pointing towards the correct answer. The critical feedback of the volunteers helped to revise some of the questions and the design of the experiment. The results of this pilot experiment are provided in Appendix C.

At the end of this experiment, the interface was developed and a second pilot experiment was performed, with the goal of testing the interface. It involved four different volunteers. The results from this pilot experiment and the feedback from its participants allowed further refinement of the questions and answers and of the web interface. After the two pilot experiments, the actual experiment was widely advertised in various NLP and Linguistics mailing lists and encountered the interest of a large number of participants. The volunteers who agreed to participate were 103. Figures 5.2, 5.3, and 5.4 show the numbers of participants per gender, level of English, age, profession and native language.

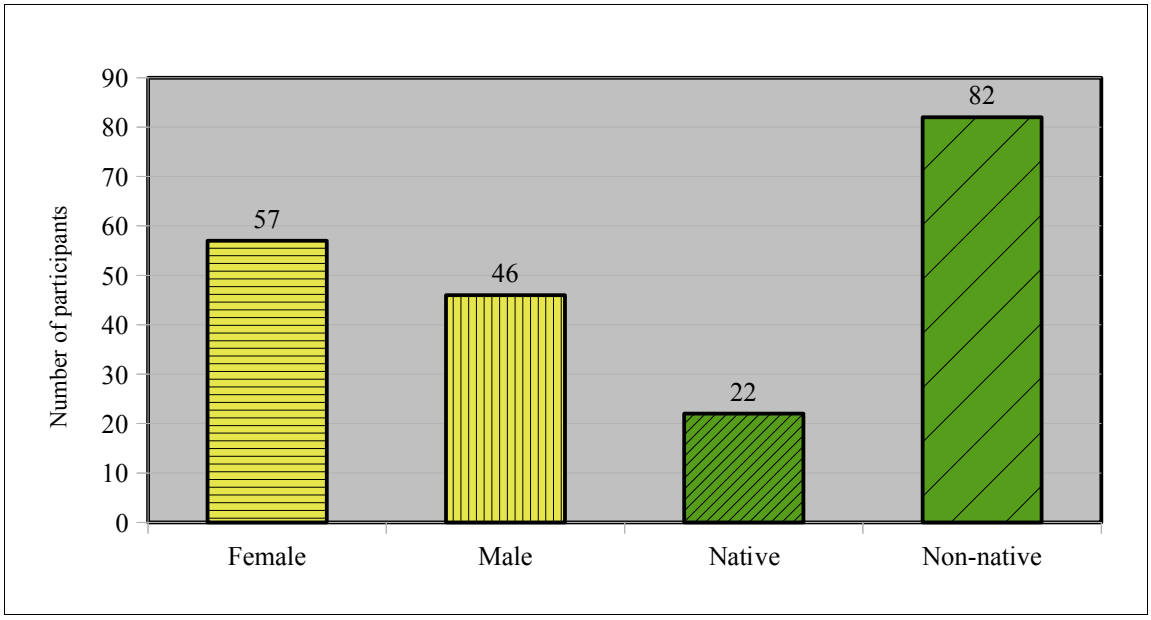


Figure 5.2.: Number of participants per gender and level of English.

It can be seen that while the proportions of Male and Female participants are similar, the difference between numbers of participants who indicated their level of English as Native and Non-native is 76%. The professions and native languages entered by participants were clustered into major groups (as explained in sections 5.4.5 and 5.4.6) and showed in Figures 5.3 and 5.4.

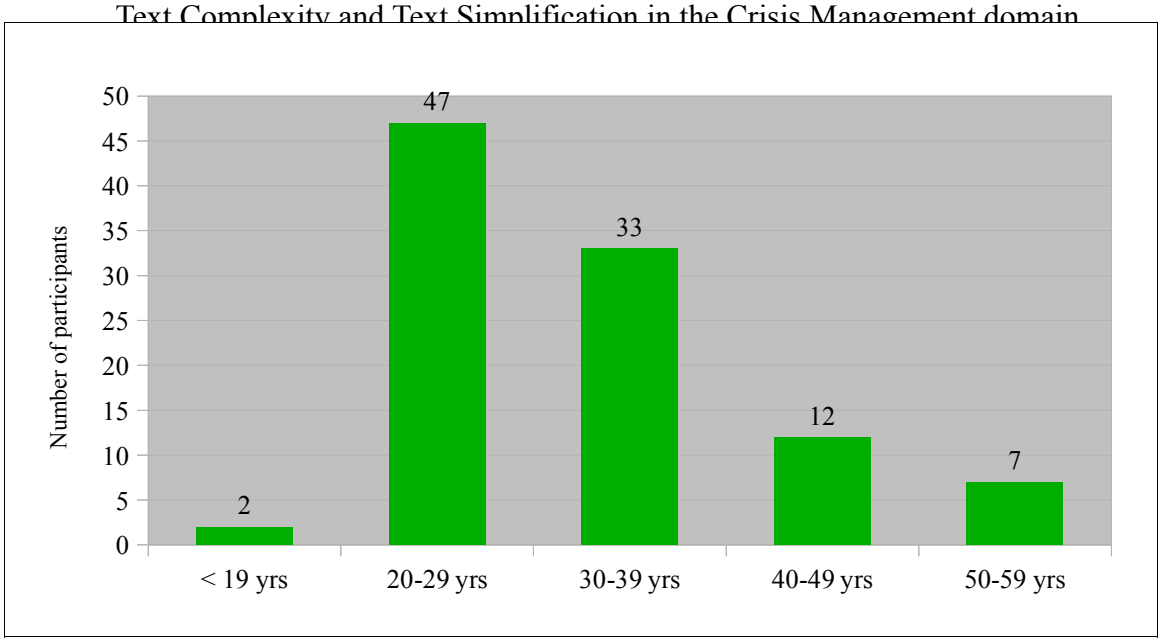


Figure 5.3: Number of participants per age.

As Figure 5.3 shows, most of the participants in the experiment are between 20 and 29 and 30 and 39 years old. The distribution of participants per profession is displayed in Figure 5.4.

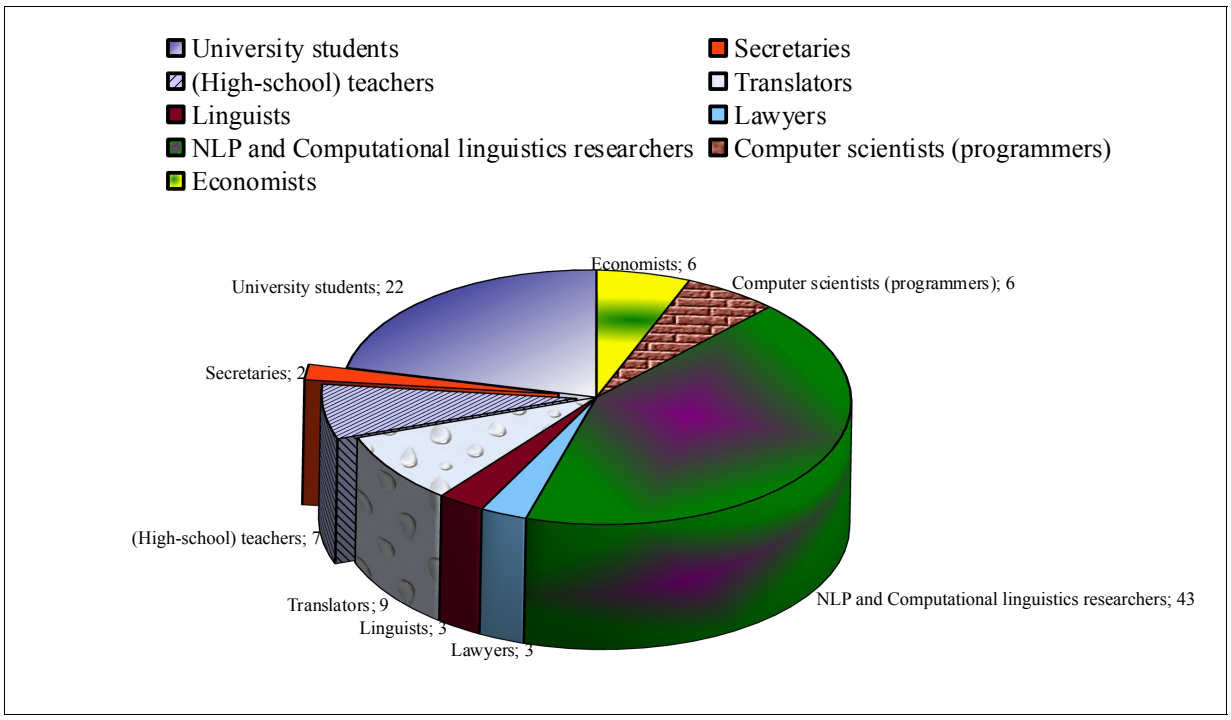


Figure 5.4: Number of participants per profession.

As in Figure 5.3, the numbers next to each section in Figure 5.4 indicate the number of participants in that group. It can be seen that among professions, the highest number of participants are University students and Natural Language Processing researchers.

Figure 5.5 shows that the largest groups of participants are those who indicated their native language to be English, Romance, and the Southern Slavic and Germanic languages. The picture also suggests that further splitting of languages in Indo-European and non-Indo-European could generate a new large group.

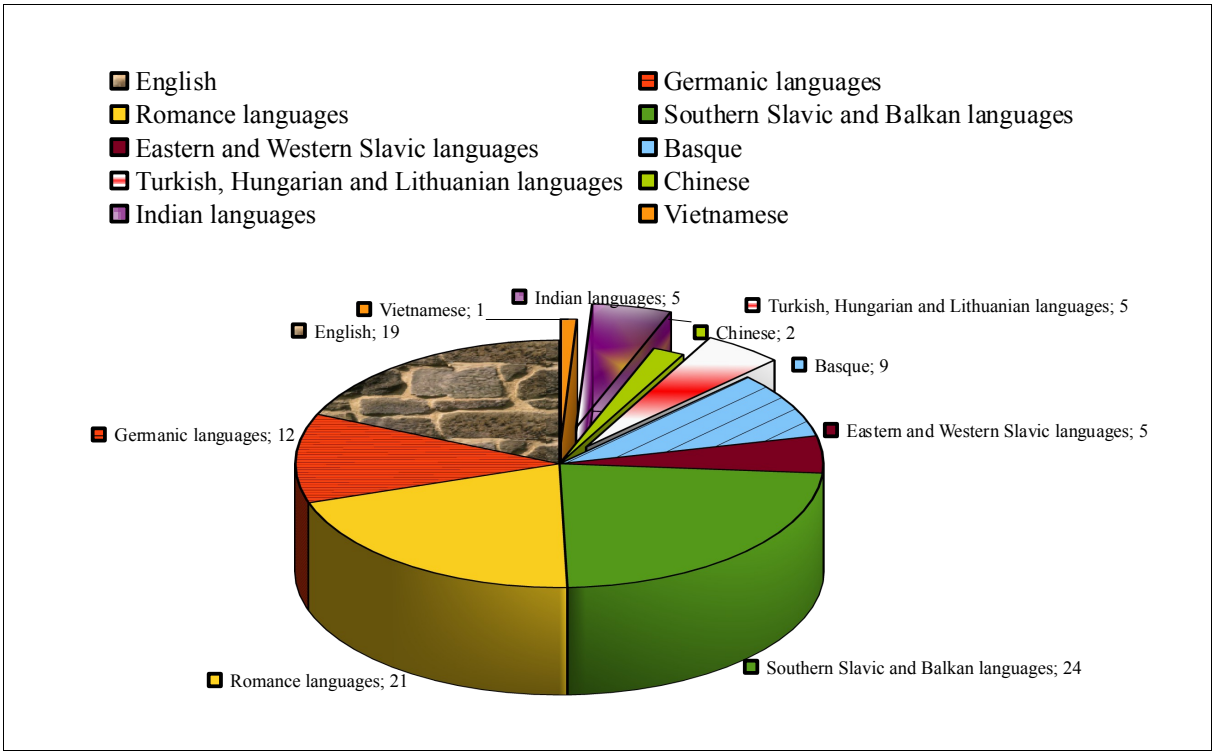


Figure 5.5: Number of participants per native language.

The time taken to prepare the experiment was about nine months. The time taken for running it with real users took around a month and a half, plus two more months of evaluating the data. Next, Section 5.4 will present the results collected during the proper experiment.

5.4. Experiment Results

As stated in Section 5.3.1, the first extrinsic evaluation of CLCM consisted of testing the research hypothesis that CLCM has a positive impact on human comprehension under stress.

As has also been explained, this research hypothesis was tested by running a large-scale online experiment involving a large number participants, who were asked to read texts in a limited amount of time and reply to the questions after them. There were two methods of testing whether CLCM has a positive impact on their reading comprehension: namely, to measure and compare time to

reply to questions, and more concretely, to provide correct answers; and to compare the proportion/percentage of correct answers to questions about the original (complex) and about the simplified text. The time assessment was restricted to measuring the time to provide correct answers, because this indicates that the participants have understood correctly the text. In the analysis of the time employed to reply correctly to questions, the time, which was originally recorded in milliseconds (as explained in Section 5.3) was divided by 60 and in this way transformed into the so-called 'milliminutes'. The testing is thus based on the following two assumptions:

1. The CLCM simplification has a positive impact on reading comprehension under imitation of a stress situation if the percentage of correct answers given to questions about the simplified text is higher than the proportion/percentage of correctly answered questions asked about the original (complex) text.

2. The CLCM simplification has a positive impact on reading comprehension under imitation of a stress situation if the time employed by the participants to identify the correct answer and answer correctly to the questions about the simplified text is smaller than the time employed by the participants to identify the correct answer and answer correctly to the questions about the original (complex) text.

This section will present the results obtained in this experiment. As the experiment has encountered a significant amount of interest from the research community and a large number of volunteers from different countries participated, and thus a large number of different users' data were collected, the results were divided into different perspectives for both time to answer the questions and proportion of correct answers:

1. A comparison of the results of **all participants** for the complex and the simplified text (Section 5.4.1).

2. A comparison of the results of the participants for complex and the simplified text **with a focus on whether they are *native* or *non-native* speakers of English** (Section 5.4.2).

3. A comparison of the results of the participants for the complex and the simplified text **with a focus on whether their gender is *female* or *male*** (Section 5.4.3).

4. A comparison of the results of the participants for the complex and the simplified text **with a focus on their *age*** (Section 5.4.4).

5. A comparison of the results of the participants for the complex and the simplified text **taking into account their *professions*** (Section 5.4.5).

6. A comparison of the results of the participants for the complex and the simplified text **taking into account their *native languages*** (Section 5.4.6).

Although information regarding the participants' experience in the Crisis Management domain was also collected, it has not been taken into consideration, because there was a very low number of participants with such experience. Next, Section 5.4.1 will present the results regarding all of the participants in the experiment as a whole.

5.4.1. Results for all participants

The total number of participants in this experiment was 103 people. The results of the time employed by the participants to identify the correct answer and answer correctly to the questions about an original (complex) or a simplified text are shown in Figure 5.6, while the proportion/percentage of correct answers given to questions about the two texts are provided in Figure 5.7.

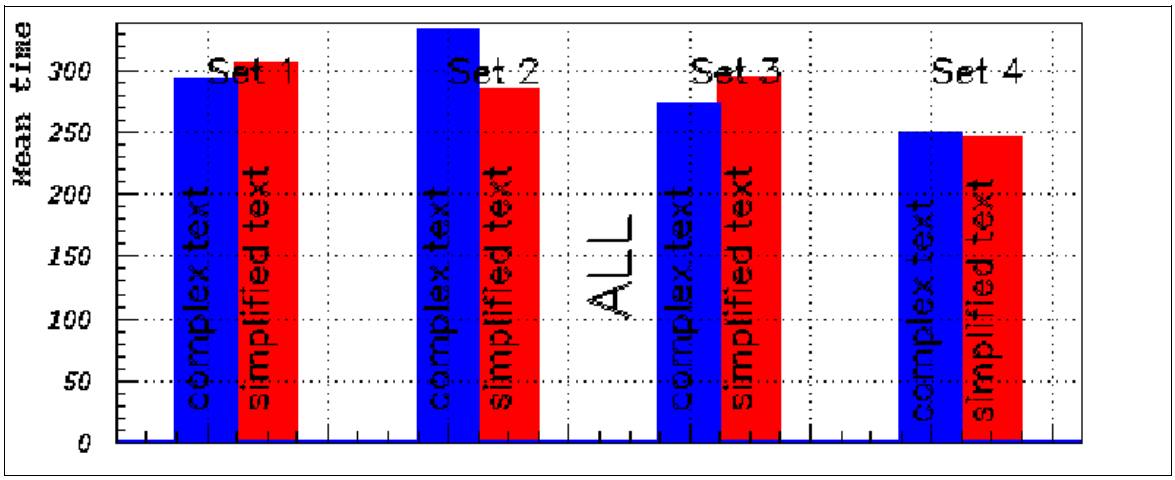


Figure 5.6: Time to correctly answer questions for all participants.

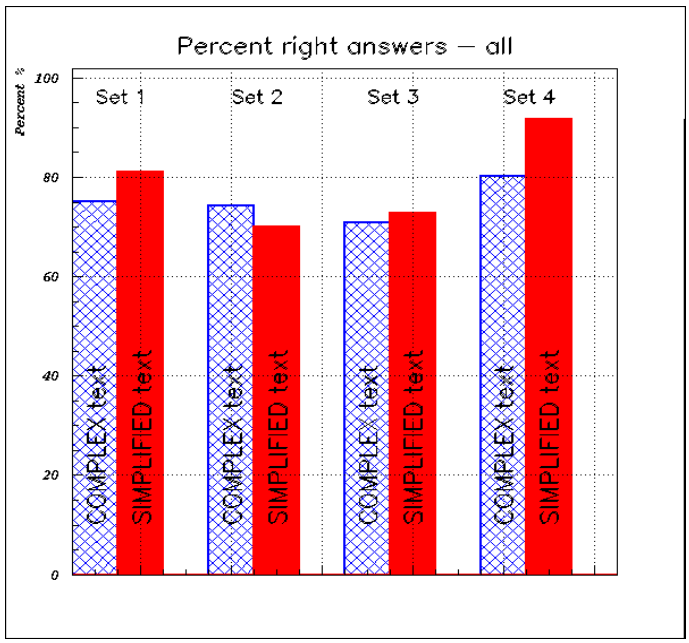


Figure 5.7: Percentage of correct answers for all participants

Both Figure 5.6 and Figure 5.7 are composed in a similar way. Both of them have the single texts divided into sets on the x axis. The sets are ordered from left to right, namely: Set 1, Set 2, Set 3, and Set 4. The y axis in Figure 5.6 contains the mean time measured in milliminutes for correctly answering the questions, divided by number of answers, while the y axis of Figure 5.7 shows the percentage of correctly answered questions. The time was normalised per number of answers because, due to the limited number of participants and to the randomisation of questions, there are a different number of participants per question. Thus, in Figure 5.6 the columns represent the average time spent to answer correctly questions, while in Figure 5.7 the columns represent the percentage of correct answers per text. In each figure, the first column of the set represents the original (complex) text and the second column reflects the simplified text results. According to the assumptions stated at the beginning of Section 5.4., if the CLCM simplification has a positive impact on reading comprehension in this experiment, then the second columns of each set should be lower in Figure 5.6 (i.e. less time employed for answering the questions about the simplified text) and higher in Figure 5.7, i.e. a higher number of correct answers was given while answering questions about the simplified text. The comparison of the time employed to answer questions correctly did not yield results pointing clearly to the positive impact of the CLCM simplification on reading comprehension. Figure 5.6 shows that the time employed to give correct answers is lower for the simplified text for Set 2 and Set 4, and slightly higher for the other two sets. These results are not statistically significant, and thus this general picture does not provide any evidence for either a positive or negative impact of CLCM on reading comprehension. In contrast, Figure 5.7 shows a higher number of correctly answered questions about the simplified text in three out of the four sets (Set 1, 3 and 4), with larger differences for Set 1 (5%) and 4 (~12%). The differences displayed in Figure 5.7 support the research hypothesis that the CLCM simplification has a positive impact on reading comprehension under stress, as assessed in terms of the percentage of correct answers. It should be noticed that Set 4 in general shows lower time for both texts and a larger number of

correct answers for both texts, which may mean that it is a text which is simpler to comprehend, or that the questions following the texts of set 4 are simpler than those following the other three sets of pairs of texts. As the results for Set 1 and 2 in Figure 5.6 are not statistically significant, but the results for Set 3 are significant with 93% confidence, and those for Set 4 are statistically significant at over 99% confidence, it is considered that this finding supports the research hypothesis.

5.4.2. Native/non native speakers of English results

The inconclusive results for all participants motivated the need to have a deeper look into the impact of CLCM on particular groups of participants, and therefore further analysis of the data by focusing on particular variables related to the groups of participants. The first variable to be analysed was whether the participants were *Native* or *Non-native* speakers of English, as in the first place, in the modern global world there are large numbers of non-native readers, whose correct understanding of emergency instructions needs to be guaranteed as well; and in the second place, non-native speakers of English are often target readers or users of the readability formulae and controlled languages presented in Chapter 2, and thus it was assumed that the non-native speakers would experience more difficulties understanding complex text than the English native speakers.

The number of native speakers in the experiment was 22, versus 82 non-native speakers. The native speakers of English came primarily from the UK and the United States. The analysis of the time employed to answer questions did not show any significant difference for participants divided on the native/non-native axis. On the other hand, the analysis of the number of correct answers for the *Non-native* participants (Figure 5.8) showed differences, which were statistically significant with around 70% confidence.

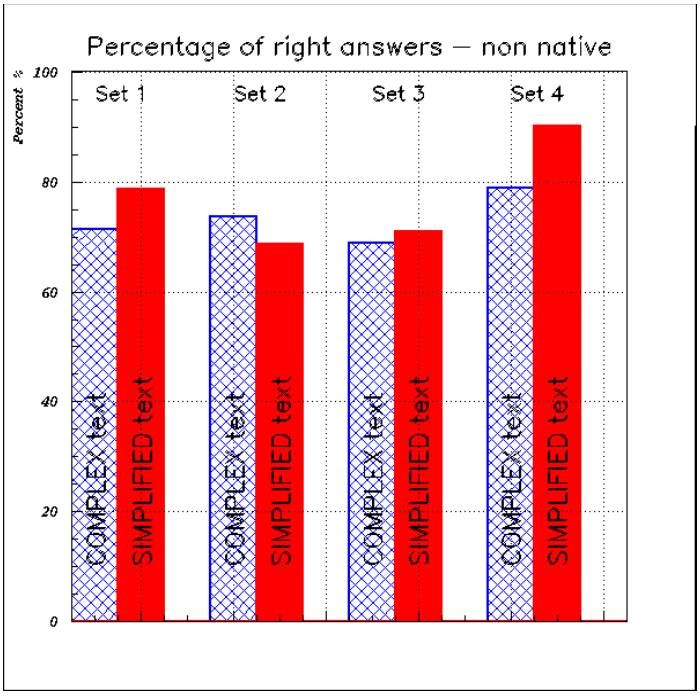


Figure 5.8: Percentage of correct answers for *Non-native* participants.

As can be seen, Figure 5.8 shows a larger number of correct answers again for Sets 1, 3, and 4. This partially supports the research hypothesis, but the statistical significance is not large. This motivated an analysis based on dividing the participants into more fine-grained groups.

5.4.3. Gender results

The inconclusiveness of the results obtained in Sections 5.4.1 and 5.4.2 motivated a further look into the impact of the CLCM simplification on more fine-grained groups of participants. After the *native/non-native* analysis, another variable—the gender of the participants—was analysed. This variable was taken into account due to the fact that there are known differences in the mental processes and particularly in decision making for different genders (Sanz de Acedo Lizarraga et al., 2007).

There were fifty-seven female and forty-six male participants. The analysis of the impact of CLCM

on reading comprehension was focused on two subdivisions of the participants:

- The simple division of participants into female and male participants (*Female* and *Male*).
- A division into female and male participants also taking into account the previous variable, namely whether they are *native* or *non-native* speakers of English.

The comparison of the time employed by the participants to identify the correct answer and answer correctly to questions for female and male participants is provided in Figure 5.9. It shows interesting results.

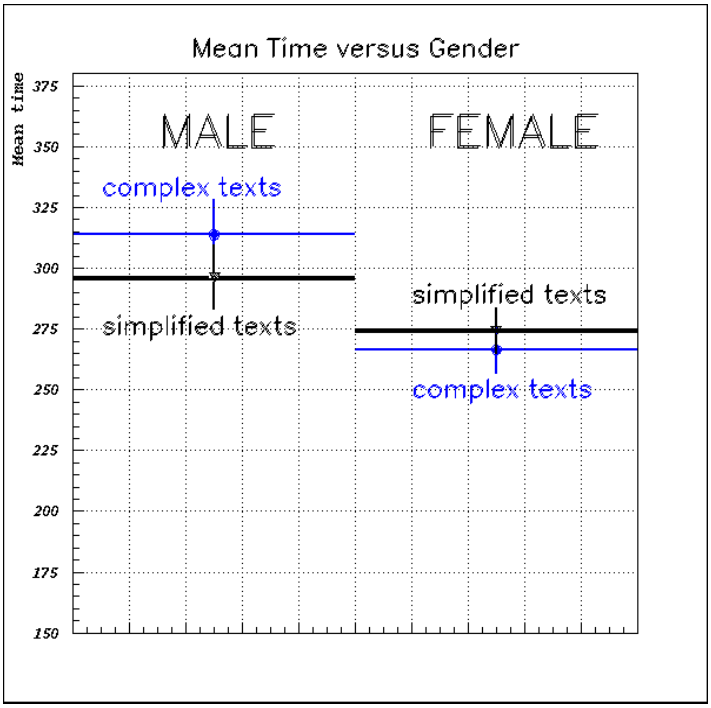


Figure 5.9: Time to correctly answer questions for *Female* and *Male*.

Figure 5.9 displays the times employed by the participants to identify the correct answer and answer correctly to questions formulated about the texts on the ***y*** axis, while the ***x*** axis displays the two groups of data—male participants and female participants. The blue line, or the line with the circle, represents the complex text, while the red line, or the line with the square, represents the simplified

text. As can be seen, in general, *Male* participants employ more time to correctly answer questions for both texts than *Female* participants. In addition, *Male* show a decrease in time to correctly answer questions about the simplified text as compared with the complex one, where they employ more time, while *Female* employ more time correctly answering questions about the simplified text. The results lead to the conclusion that the CLCM simplification has a positive impact on reading comprehension of *Male* participants, but a slightly negative effect on the reading comprehension of *Female* participants. The differences between female and male are small (p=0.13 for *Male* and p=0.18 for *Female*). The differences between *Male* and *Female* for the same text are significant at p<0.01 for the complex text and at p=0.16 for the simplified text. The increased mean times to give correct answers for the simplified texts for *Female* can also be seen in Figure 5.10.

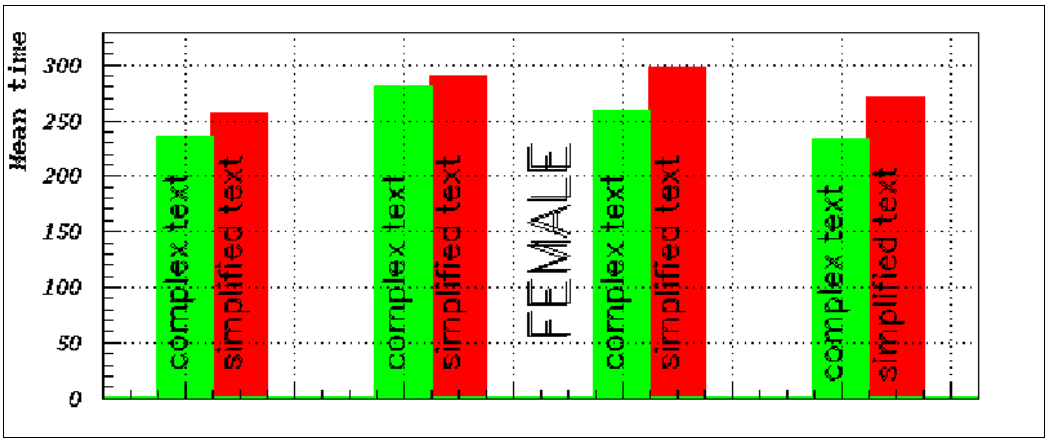


Figure 5.10: Time to correctly answer questions for *Female*.

The texts in Figure 5.10 are again presented as pairs of complex-simplified text and are given in the order Set 1, Set 2, Set 3, and Set 4. As can be seen from Figure 5.10, there is a clear increase for all simplified texts in the time employed by the participants to identify the correct answer and answer correctly to questions, as the columns at even positions are higher. The lesser time that female participants employ in giving correct answers may be explained by women's better reading skills (Lietz P., 2006), while the poor impact of text simplification on *Female* participants could be explained by the discoveries of some studies that women collect more information about the environment before making a decision (Sanz de Acedo Lizarraga et al., 2007). As the effect of the

CLCM text simplification is to produce shorter sentences and thus reduce context, this could hinder women's comprehension. The analysis of the percentage of correctly answered questions by gender showed the best results for *Female*, and particularly for *Female non-native* speakers, as can be seen in Figure 5.11.

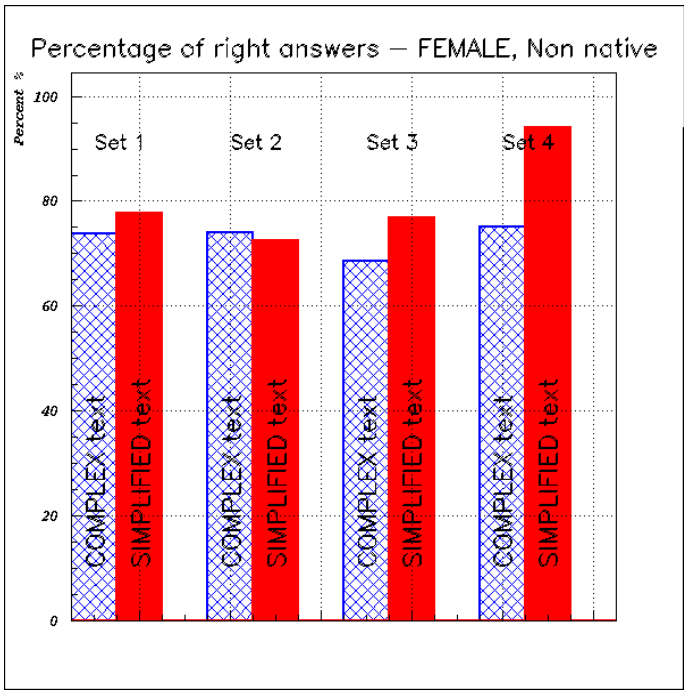


Figure 5.11: Percentage of correct answers for *Female non-native* speakers.

As can be seen in Figure 5.11, in three out of the four sets of texts, the percentage of correct answers is higher for the simplified text. The differences between the proportions in Set 1 and Set 2 are not statistically significant, but for Set 3 they are statistically significant with 91% confidence, and for Set 4 they are statistically significant with over 99% confidence. The results in Sets 3 and 4 mean that the CLCM simplification has a positive impact on *Female non-native* speakers, which supports the research hypothesis.

The same analysis run for male participants shows that according to the percentage of correct answers, the group that most benefits from the CLCM simplification among *Male* is *Male Native*. This can be seen in Figure 5.12.

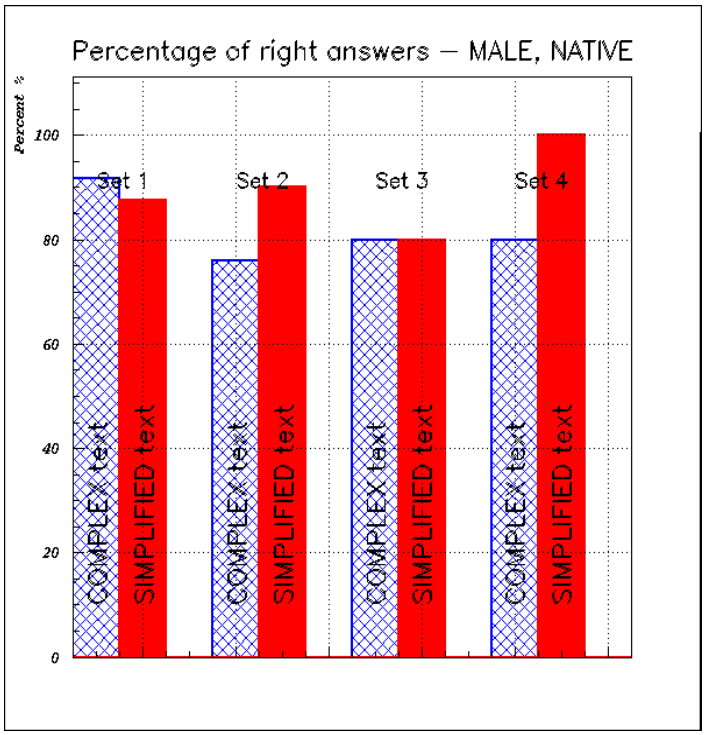


Figure 5.12: Percentage of correct answers for *Male Native*.

As can be seen in Figure 5.12, there is a much higher proportion of correct answers for Set 2 and Set 4, while Set 1 and Set 3 do not show positive results for the CLCM simplification. Particularly, in Set 4, the percentage of correct answers for the simplified text is 100%, i.e. all questions were answered correctly and the difference with the complex text for the same set is as much as 20%. However, it should be noted that the number of native speaker male participants in the experiment was very limited—there were only seven—so it would be difficult to interpret statistical significance.

5.4.4. Age results

Similarly to gender, the age of the participants was analysed, as there is a known decline in reading comprehension, working memory performance, and reading rates for older ages (Norman, 1991, Sanz de Acedo Lizarraga et al., 2007). The age of the participants in this experiment varied from 18

years old to 54 years old. While the analysis of the percentage of correct answers did not yield revealing results, the analysis of the time employed by the participants to identify the correct answer and answer correctly to questions produced interesting results. These results can be seen in Figure 5.13.

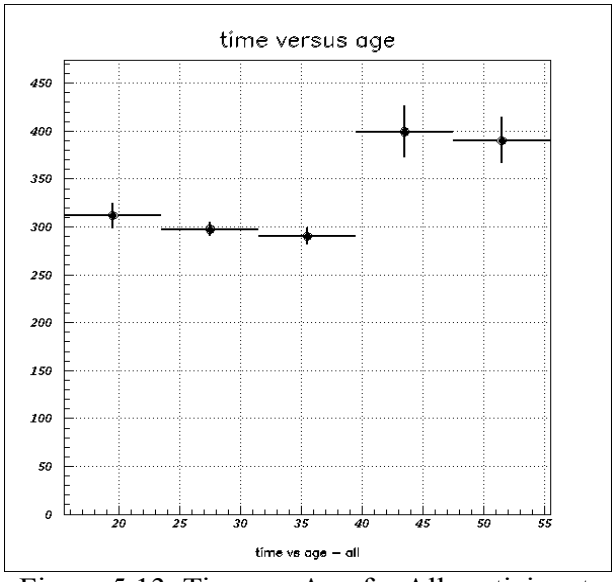


Figure 5.13: Time vs. Age for All participants.

As can be seen in Figure 5.13, the x axis contains the ages of the participants, with the results divided into five age groups, while the y axis contains the time to correctly answer the questions about all texts (both complex and simplified). The purpose of Figure 5.13 is to analyse the distribution of times according to age groups. The dots on the horizontal lines represent the mean times to give correct answers. As can be seen, while the age groups below 40 years old have more or less the same mean times to answer (around 310 milliminutes), the participants of over 40 years of age have much higher mean times (around 400 milliminutes). This anticipated finding motivated further analysis of the participants' performance divided in two age groups, but the results were not statistically significant.

5.4.5. Profession results

Although the participants' background in Crisis Management has not been taken into account, due to the low number of participants with such background (4), it was considered important to investigate the impact of the CLCM simplification on the reading comprehension of participants with different professions, as it was hypothesized that some professions may have better reading skills (e.g. teachers, translators, secretaries, and linguists) than others. With the aim to create larger groups of participants in the attempt to obtain statistically significant data, the professions which the participants indicated were normalised and clustered into groups. The following list exemplifies the normalised categories of professions:

1. *High-school students* (later included in *Students)*
2. *Students* (including undergraduate, Master's and 1st year Ph.D. students)
3. *Secretaries*
4. *Teachers* (school teachers)
5. *Translators*
6. *Linguists*
7. *Lawyers*
8. *NLP researchers*
9. *Computer scientists*
10. *Economists*

The time and the percentage of correct answers per text were then calculated for each of these categories, without taking any other user variables into account (gender, native speaker status, or

age).

The analysis of the time employed to give correct answers showed good results only for the category S*tudents*. Their results are shown in Figure 5.14.

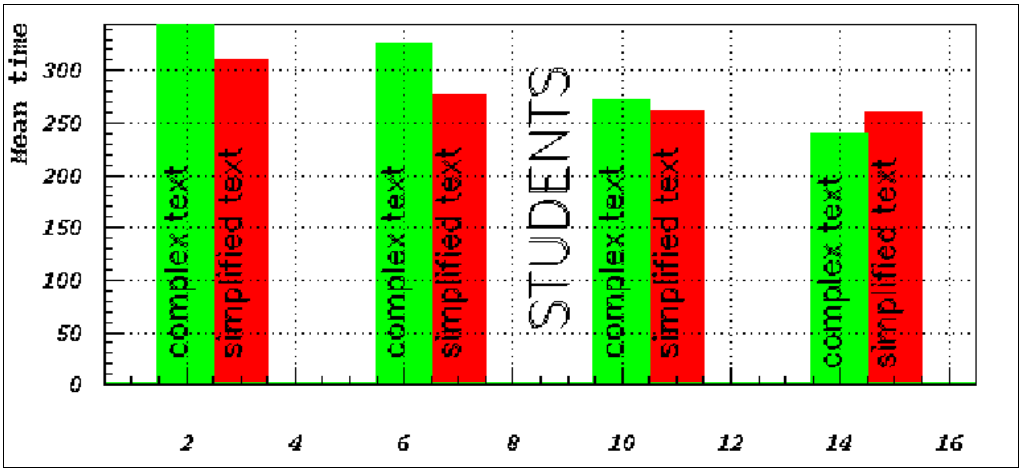


Figure 5.14: Time to correctly answer questions by *Students*.

As can be seen in Figure 5.14, there is a well-expressed positive impact of CLCM in Sets 1, 2 and 3 (differences between the means of the complex and simplified texts between 10 and 50 milliminutes), compared to Set 4, where there is a slightly negative impact (differences of about 20 milliminutes). It can be concluded that there is a clear improvement in comparison with Figure 5.6 (time for *All* participants) and as the results are significant with 87% confidence, that these differences support the research hypothesis. It was estimated that at least 34 students are necessary, in order to obtain 95% confidence. Currently there are 24 Students. The best results for the percentage of correct answers are given by the *NLP researchers* and the cluster *Translators, Linguists and Lawyers*, as can be seen in Figures 5.15 and 5.16, respectively.

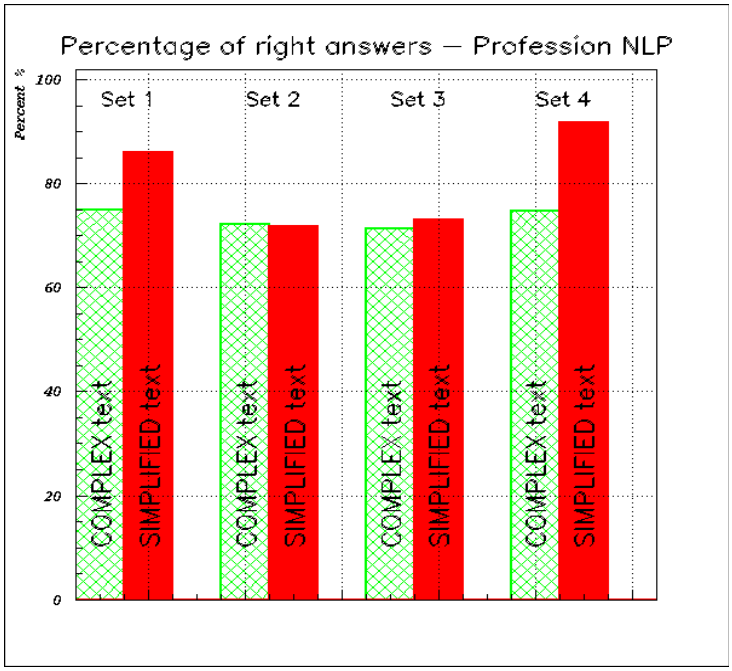


Figure 5.15: Percentage of correct answers for *NLP researchers*.

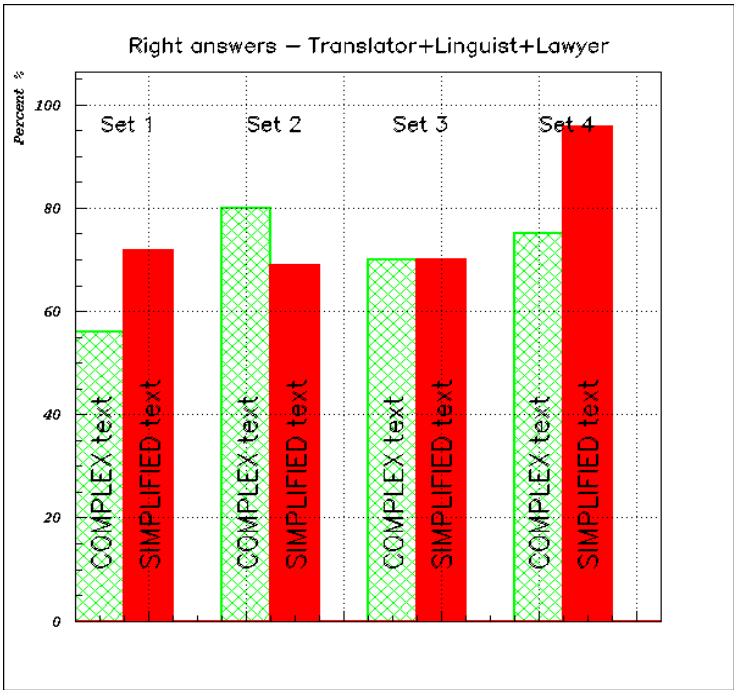


Figure 5.16: Percentage of correct answers for *Translators, Linguists and Lawyers*.

As can be seen, the figures for *NLP researchers* can be considered better than those for *Translators, Linguists and Lawyers*, with positive differences between the means in Figure 5.15 for *NLP researchers* of 1-12% for Sets 1, 3, and 4, and a negative difference of 1% in Set 2, although the

positive differences between the means in Figure 5.16 for *Translators, Linguists and Lawyers* are about 11-16% for Sets 1 and 4, but there is no difference for Set 3 and a negative difference of 10% for Set 2. The statistical significance for *NLP researchers* is 96% for Set 1, not significant for Set 2 and Set 3, and significant at over 99% confidence for Set 4, which supports the research hypothesis for the category of *NLP researchers*. The results for *Translators, Linguists and Lawyers* are not significant for any of the sets, which can be explained by the lower number of professionals from these categories (9 translators, 3 linguists and 3 lawyers).

In comparison, the high statistical significance of the NLP profession is also due to the fact that this is the largest profession group, as can be seen in Figure 5.17. The distribution of users per texts according to their profession can be seen in Figure 5.18. The horizontal axis shows the profession categories, while the vertical axis contains the numbers of answers.

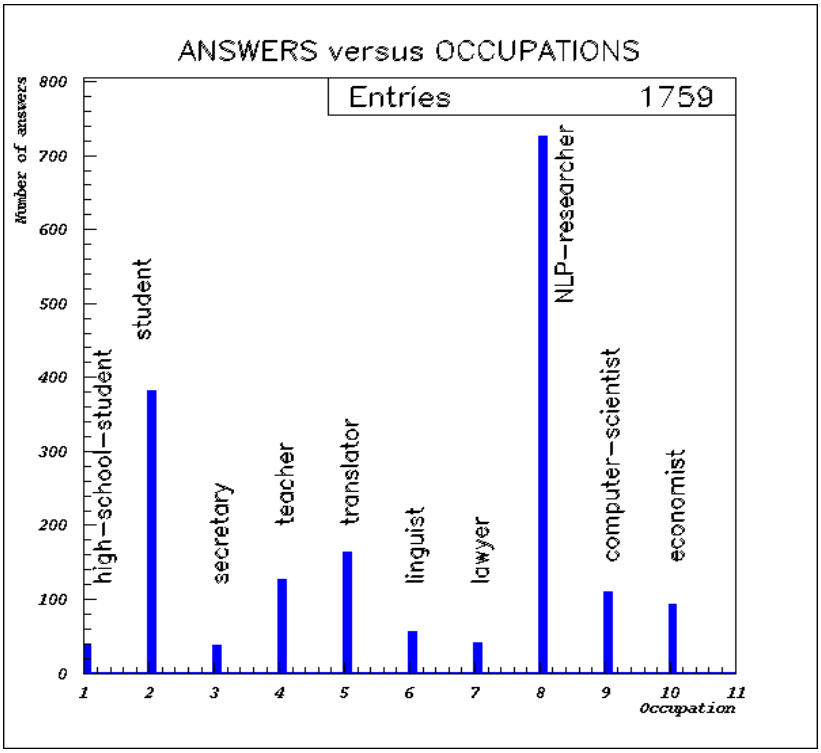


Figure 5.17: Number of answers per profession.

As can be seen in Figure 5.17, the distribution of answers per profession is not uniform, with *NLP researchers* being the largest group and *High-school students* the smallest. The number of

participants per profession can be deduced from this graphic by dividing the number of answers (vertical columns) by approximately nineteen units (number of answers per person), because, as explained in Section 5.3.3.6., the data is recorded answer per answer and each participant replies to nineteen questions (4 or 5 for each of the four texts). In this way can be seen that the number of *High-school students* is 2. The small number of participants for certain professions motivated the need to aggregate them in larger groups, as was done for the group of *Translators, Linguists and Lawyers*. Figure 5.18 shows the distributions of participants per text, with the horizontal axis indicating the text number and the vertical – the number of participants.

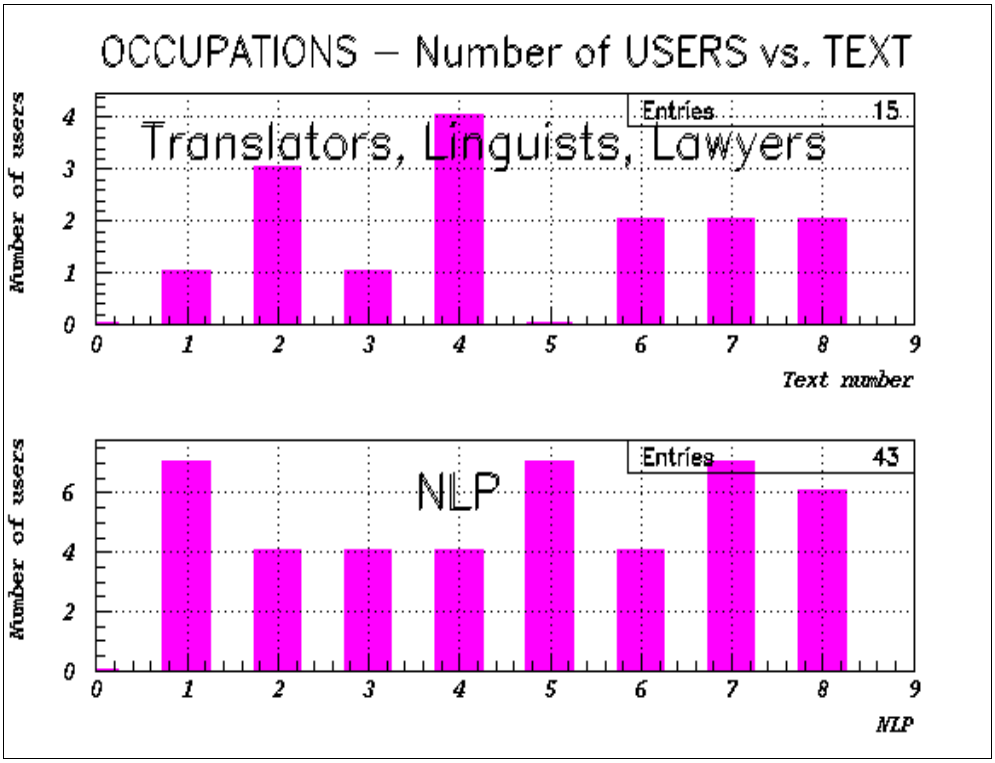


Figure 5.18: Distribution of participants per text according to profession.

As can be seen in Figure 5.18, the number of *Translators, Linguists and Lawyers* per text vary between one and four, while for *NLP-researchers* – between four and eight participants.

Due to the low number of participants from certain professions and to the displaying of a random combination of two simple and two complex texts to every single participant, the distribution of

users per text is also not even, as can be seen in Figure 5.18. As can be seen in Figure 5.18, the group of *Translators, Linguists and Lawyers* do not actually cover all texts, while the *NLP researchers* do. Next, Section 5.4.5 will provide the final analysis of the participants' data.

5.4.6. Native language results

The last participant variable taken into consideration, related to the subdivision of into *Native/Non-native* speakers analysed in Section 5.4.2, is the native language of the participants. This enabled the further examination of the impact of CLCM on different kinds of *Non-native* English speakers. The rationale for taking this variable into consideration was based on the assumption that participants who are native speakers of languages very different from English (e.g. Japanese) may have bigger problems with reading comprehension than native speakers of languages closer to English (e.g. Dutch or Spanish). In the attempt to create larger groups of participants, the native languages entered were normalised and clustered into categories on the basis of the greatest similarity between them, similarly to the professions. The main criteria on which the language grouping was based on were language families (Fromkin and Rodman, 1978) and geographical proximity. The list of categories obtained follows below:

1. *English language* (same as *Native* speakers)
2. *Germanic languages* (German and Dutch)
3. *Romance languages* (French, Spanish, Italian, Portuguese, Catalan)
4. *Southern Slavic and Balkan languages* (Bulgarian, Greek, Serbian, Croatian). Greek was included in this group, because there was only one participant with Greek and Greek could not be grouped with any other Indo-European languages, because it belongs to an independent branch of the Indo-European languages (Hellenic) (Fromkin and Rodman,

1978). For this reason, it was grouped together with the Southern Slavic languages for geographical reasons.

5. *Eastern and Western Slavic languages* (Russian and Czech)
6. *Basque language*
7. *Turkish, Hungarian and Lithuanian languages.*
8. *Chinese languages*
9. *Indian languages* (Bengali, Oriya, Kannada and Malayalam). Although two of the languages in this group are from the Indo-Iranian family (Bengali and Oriya) and the rest are from the Dravidian family (Kannada and Malayalam), these languages were grouped together on the basis of geographical proximity and similar educational system. It was considered that due to this fact the native speakers of these languages are likely to have similar proficiency in English as a result.
10. *Vietnamese language*

Figure 5.19 shows the distribution of answers per language category, ordered according to the category numbers in the list above. The horizontal axis shows the categories of native languages, while the vertical axis contains the numbers of answers.

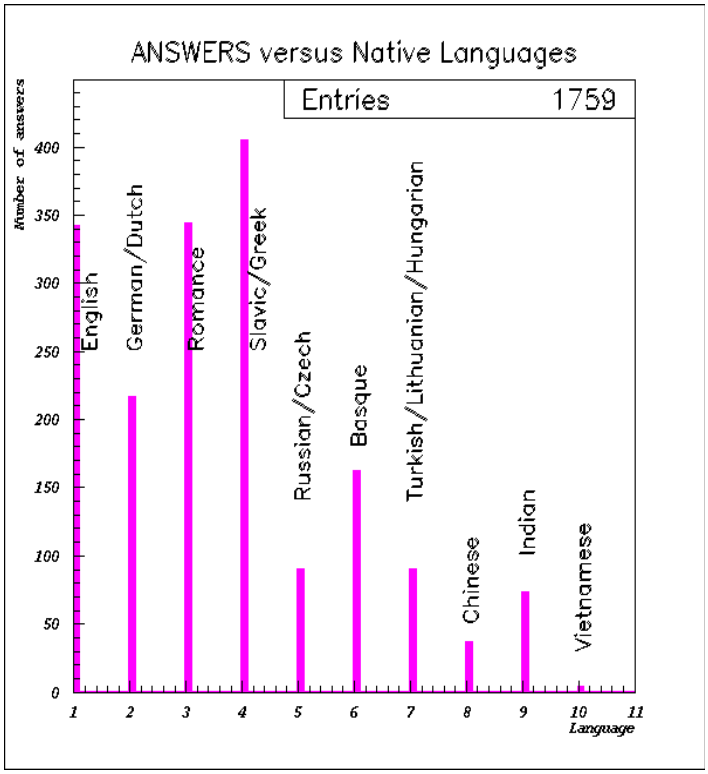


Figure 5.19: Distribution of Answers per Native Languages.

As can be seen from Figure 5.19, the distribution of answers per language group is not uniform, with the most answers (and thus number of participants) being the native speakers of *English language*, *Germanic languages*, *Romance languages,* and *Southern Slavic and Balkan languages*.

The best results showing the impact of the CLCM simplification on reading comprehension of participants with different native languages for time to give correct answers are those of the following clusters:

8. Chinese languages + 9. Indian languages

6. Basque language + 7. Turkish, Hungarian and Lithuanian languages

The results per text regarding the time employed by the participants to identify the correct answer

and answer correctly to questions are shown in Figure 5.20.

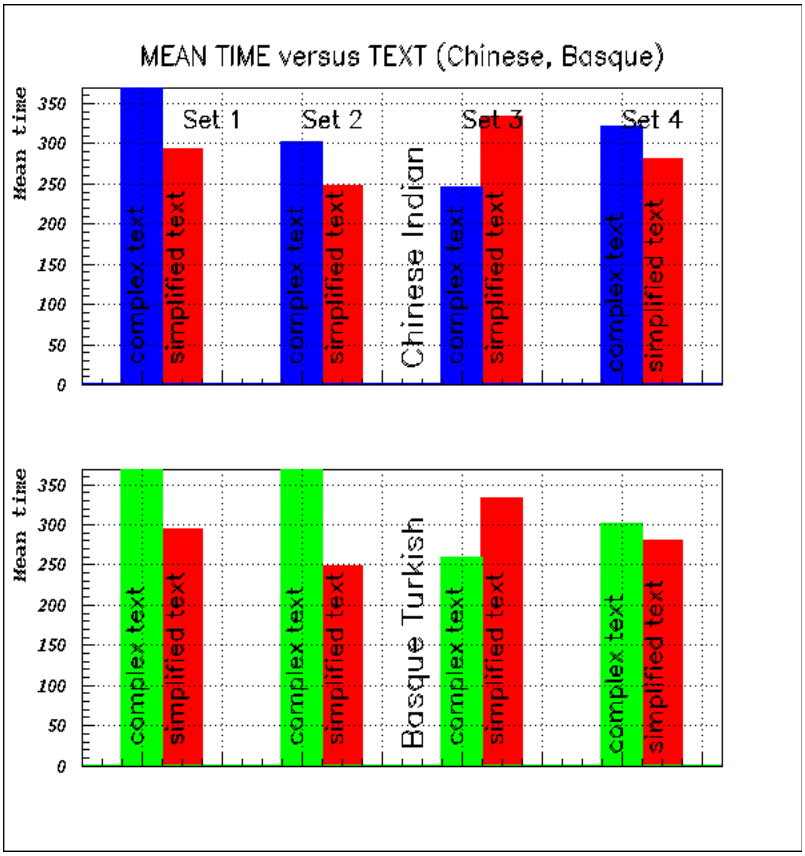


Figure 5.20: Time to correctly answer questions by Clusters 8+9 and 6+7.

As can be seen from Figure 5.20, both clusters exhibit positive results for Sets 1, 2, and 4, and negative results for Set 3. These results (three out of four sets with a positive impact) support the research hypothesis.

The best results for this variable regarding the percentage of correct answers are those of the *Basque language* (which was allowed to be placed in a separate category due to the large number of participants who were speakers of this language, as can be seen from Figure 5.19) and of the cluster *8. Chinese languages + 9. Indian languages + 10. Vietnamese language*. Their results can be seen in Figures 17 and 18, respectively.

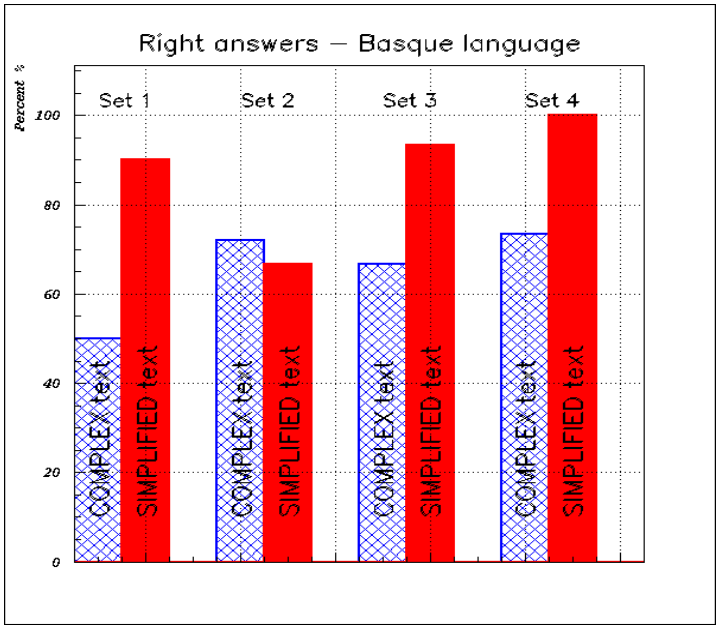


Figure 5.21: Percentage of correct answers of native speakers of the Basque language.

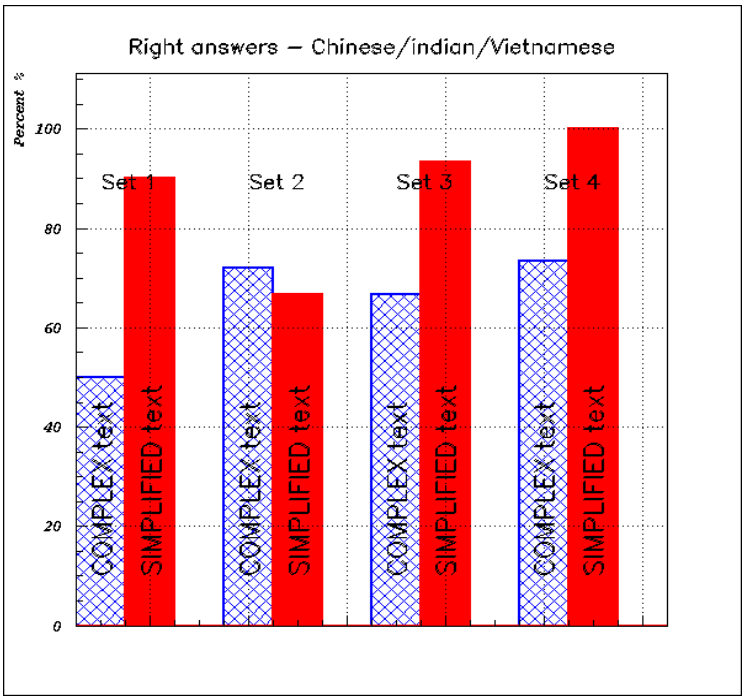


Figure 5.22: Percentage of correct answers of native speakers of the Cluster 8+9+10.

As can be seen from Figures 17 and 18, the situation with these native speaker language groups is similar— Sets 1, 3, and 4 are very positively influenced by simplification (differences between the proportions between 17% and 40%), while Set 2 is negatively influenced, with a small difference between the proportions (6%). Particular attention needs to be payed to Set 4 in both Figures 17 and 18, in which the percentage of correct answers for the simplified text is 100%. As the participants with these native languages were very low in number (nine for Basque and fewer for the others), statistical significance was not calculated. The above results support the research hypothesis investigated. The problems with Set 2, which were seen many times before, suggest that there is a clear problem with this set of texts.

Finally, similarly to the case of professions and due to the different number of participants per native language and the randomisation of displayed texts, the number of native languages per text was not equal, as can be seen in Figure 5.23. The vertical axis in Figure 5.23 contains the number of users per text, while the horizontal axis shows the texts from 1 to 8.

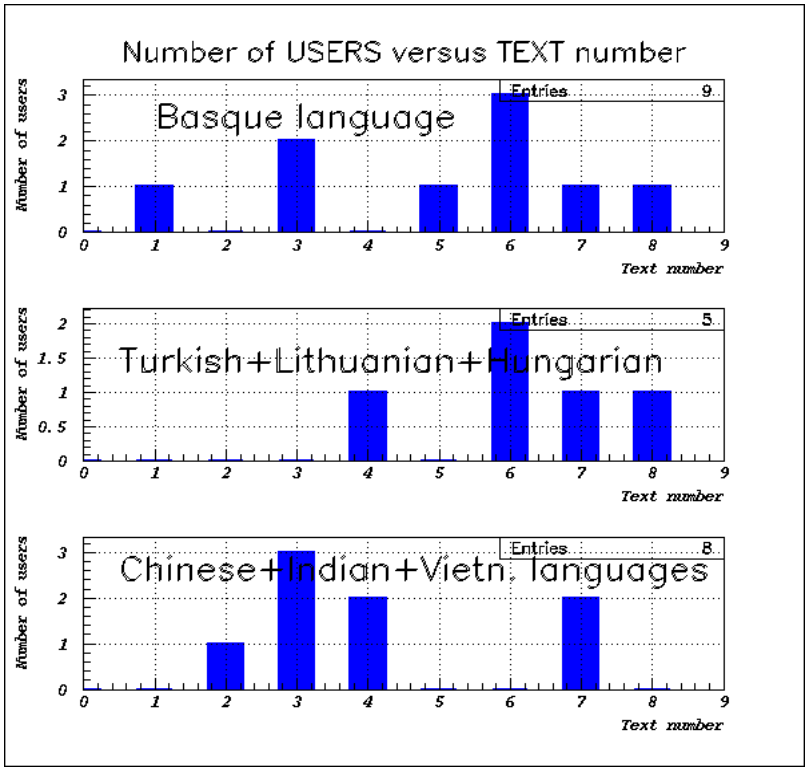


Figure 5.23: Distribution of participants per text for Categories 6, 7, 8, 9, and 10.

As can be seen, these groups are very small: nine users in total for *Basque* and five users for *Chinese, Indian and Vietnamese languages*. The distribution of participants per text is not uniform, especially with so few participants for these languages (between zero and three participants per text). Next, Section 5.5 will provide a summary and discussion of these findings.

5.5. Summary of the Findings and Discussion of the Results

This section will present a summary of the findings presented in Section 5.4 and their discussion. The results presented in Section 5.4 aimed to test the research hypothesis that the CLCM simplification has a positive impact on reading comprehension, which was investigated by studying two factors recorded in the experiment, namely:

1. The CLCM simplification has a positive impact on reading comprehension under imitation of a stress situation if the percentage of correct answers given to questions about the simplified text is higher than the proportion/percentage of correctly answered questions asked about the original (complex) text.

2. The CLCM simplification has a positive impact on reading comprehension under imitation of a stress situation if the time employed by the participants to identify the correct answer and answer correctly to the questions about the simplified text is smaller than the time employed by the participants to identify the correct answer and answer correctly to the questions about the original (complex) text.

Section 5.5.1 will summarize the findings relative to the reading comprehension performance of specific groups of participants, while Section 5.5.2 will provide a critique of the experiment and

directions for future work.

5.5.1. Findings regarding particular groups of participants

As was seen in Section 5.4.1, the comparison of the times employed to provide correct answers for the complex and simplified text did not give any clear indications of whether the CLCM simplification has a positive or a negative impact on reading comprehension, while the comparison of the percentage of correct answers supports the research hypothesis.

The analysis of particular groups of participants with a focus on particular user variables has lead to the following findings:

1. The CLCM simplification has a positive impact on particular gender, age, profession and native language groups of participants, namely:

    1. The CLCM simplification has a positive impact on the reading comprehension of *Male* participants.

    2. The CLCM simplification has an adverse effect on the reading comprehension of *Female* participants. This can be explained by the fact that the simplification leads to less context.

    3. The CLCM simplification has a positive impact on the reading comprehension of *Female non-native* speakers of English.

4. The CLCM simplification has a very positive impact on the reading comprehension of *Male native* speakers of English.

5. The CLCM simplification has a positive impact on the reading comprehension of *Students*.

6. The CLCM simplification has a positive impact on the reading comprehension of *NLP researchers*.

7. The CLCM simplification has a positive impact on the reading comprehension of participants who are native speakers of non-Indo-European languages, such as *Basque, Chinese, Indian,* and *Vietnamese*.

2. Some general findings about the reading comprehension of particular groups of participants were discovered, namely that:

1. *Male* participants employ substantially more time to provide correct answers for both texts than *Female* participants.

2. The participants with *Age* > *40* take a longer time to provide correct answers for both texts than participants with *Age* < *40*.

5.5.2. Critique of the experiment and future work

This Section aims at presenting the criticisms which were discovered during the experiment and outline some future work. Section 5.5.2.1. will present some general observations, while Section 5.5.2.2. an original idea about a unique "comprehension factor".

5.5.2.1. General observations

Due to the complexity of the experiment, it was discovered that it had some limitations and that future refinements are desirable. Some of the limitations discovered through the analysis of the data, from participants' feedback, and on the basis of discussions with different experts follow below:

- As was seen in Section 5.4.5, and due to the method of advertising the experiment, described in Section 5.3.3, most of the participants in the experiment are from the same social level and reading skills (University students, University graduates and over). In order to have a more representative picture of the impact of CLCM on reading comprehension, a more diversified and stratified sample of participants is needed. Future work should include social groups with lower literacy levels, lower reading skills, and less general knowledge, in order to test in a more appropriate way whether the proposed text simplification method is successful.

- As can be seen in Appendix C, the simplified texts have much larger visual length on the page than the complex ones, which are more compact. This issue, and particularly the fact that due to the visual length on the page of the simplified texts, there was not enough time to

finish reading them, was mentioned in participants' feedback, and needs to be taken into account in future work in order to ensure more precise evaluation.

- Through the analysis, and as has also been seen in the results presented in Section 5.4, it was noted that the CLCM simplification had a negative impact on Set 2 and a substantially positive impact on Set 4. This particular behaviour of the texts in Set 2 needs to be further examined in future work. For the moment it was hypothesized that the topic of Set 2 (precautions when returning home after a flood) is more familiar to the general reader than the other texts, and that this may decrease the impact of text simplification.

- In the experimental set-up, the time to answer a question includes:

    1. Reading the question
    2. Reading the answers
    3. Thinking
    4. Eventually re-reading the question and some of the answers
    5. Moving the mouse in order to mark the answer
    6. Moving the mouse to click on "Next"

While operations 3, 5, and 6 depend exclusively on factors which are outside of control, operations 1, 2, and 4, in addition to external factors, such as the light and the participant's reading skills, depend also on factors whose impact can be measured, such as the characteristics of the text, composing the question, and the answers to it.

Due to the different length and complexity of questions and proposed answers, future evaluation work may include obtaining more precise calculation by dividing the time to provide correct answers by the length of the individual questions. It is suspected that the length and complexity of the question and of the answers need to be taken into account, as it is assumed that longer and more complex questions and answers will take greater time to be read and understood, and these factors may affect the time to provide correct answers. An example of questions and answers, characterized by different length and complexity, is provided in Table 5.3.

	Question 36 (Set 2)	**Question 50 (Set 4)**
Question text	According to the text, you should return to the house:	According to the text, why do you have to close windows, doors, fireplace and woodstove dampers?
Question length	Words: 10. Characters: 54	Words: 16. Characters: 96
Answers text	1. If you are told it is safe to do so. 2. After calling the fire department. 3. To avoid fire, electrocution or explosions. 4. If you smell gas.	1. To help keep ash and gases from getting into your house. Exposure to ash can harm your health. 2. To protect you while you are outdoors or while you are cleaning up ash which has gotten indoors. 3. To help you to pay attention to warnings, and to obey instructions from local authorities. 4. To help you to listen to local news updates for information about air quality, drinking water, and roads.
Answers length	Words: 27. Characters: 134	Words: 69. Characters: 384
Total length	Words: 37. Characters: 188	Words: 85. Characters: 480

Table 5.3: Comparison of questions and answers of different length

As can be seen in Table 5.3, the second column contains the first question and its answers and the

third column contains the second question and its answers. It can be seen clearly that both the second question and its answers (Question 50) are much longer, and also more complex, than those of Question 36. Taking into account these factors may make evaluation more precise.

5.5.2.2. The C-factor

As seen in the previous sections, the evaluation of reading comprehension was based on two criteria, namely on the percentage of correct answers (Pr) and on the time employed to provide the correct answers (T), and thus testing the research hypothesis that CLCM has a positive impact on reading comprehension followed two assumptions:

1. The percentage of correct answers given for the simplified text (Pr_s) will be higher than the percentage of correct answers given for the complex text (Pr_c) , i.e. Pr_s > Pr_c.

2. The time to recognize the correct answer and reply correctly to the questions about the simplified text (Ts) will be less than the time to recognize the correct answer and reply correctly to the questions about the complex text (Tc), i.e. Ts < Tc.

An idea for future work is thus to combine both measures into one and obtain a unique reading comprehension factor (C), which would depend directly on the proportion of correct answers and inversely on the mean time to provide correct answers. The way to calculate C per text is shown in formula 5.1.

(5.1.) $$C = \frac{Pr}{T_{mean}}$$

On the basis of this formula, the lower the C-factor, the worse the reading comprehension is, and

the higher the C-factor, the better the reading comprehension is. As an example, the values for all texts of the participants who were native speakers of *Basque*, *Turkish*, *Chinese*, *Indian* and *Vietnamese languages* can be seen in Table 5.4.

Set 1	
Text 1 – complex	**Text 2 - simplified**
12.7	21.0
Set 2	
Text 3 – complex	**Text 4 - simplified**
15.1	23.0
Set 3	
Text 5 – complex	**Text 6 - simplified**
26.0	16.0
Set 4	
Text 7 – complex	**Text 8 - simplified**
23.0	30.0

Table 5.4: C-factor values for *Basque*, *Turkish*, *Chinese*, *Indian* and *Vietnamese languages.*

As can be seen in Table 5.4, the first column lists the results for the complex texts, while the second one gives the results for the simplified texts. According to the hypothesis stated in this chapter, if the CLCM has a positive impact on reading comprehension, then the simplified texts should have a higher reading comprehension score than the complex ones, and thus the C-factor values in the second column should be higher than the corresponding value in the first column for each pair of

texts. As can be seen, this is true for Sets 1, 2, and 4, while the inverse holds for Set 3, which has also been seen for the times earlier.

Next, Section 5.6. will present the conclusions of this Chapter.

5.6. Conclusions

The present chapter is the first of the three chapters (5, 6 and 7) that aim to evaluate the controlled language CLCM. The evaluation described in this chapter was the assessment of CLCM in terms of reading comprehension under stress. It consisted of a large-scale online reading comprehension experiment, which employed one hundred and four users, four complex and four simplified texts, and a complex system of questions and answers.

The evaluation was focussed on testing the research hypothesis that CLCM has a positive impact on reading comprehension under stress. It was based on the two assumptions that if CLCM has a positive impact on reading comprehension, the percentage of correct answers given to the questions after reading the simplified text would be higher than the percentage of correct answers given to the questions after reading the complex text, and that if CLCM has a positive impact on reading comprehension, the average time for giving a correct answer would be lower for the questions following the simplified text than for the questions following the complex text.

Due to the large number of participants and the large variation in their characteristics (based on variables such as gender, age, level of English, professions, and native languages), it was possible both to evaluate the participants' performance as a whole and to divide them into different groups

on the basis of different user variables.

The results, detailed in Section 5.4, demonstrate that the data for the average of all of the participants show partially positive results in favour of the CLCM text simplification, and that several groups are particularly favoured (*Male*, especially *Male Native*, *Female Non-native*, *NLP-researchers*, *Students*, participants with native languages such as *Basque*, *Chinese*, *Indian* and *Vietnamese*), while other groups are partially disfavoured (such as *Female* as a whole). The results have also lead to some interesting findings in terms of general reading comprehension for particular groups of readers, for example some gender differences between *Male* and *Female* participants, as well as *Age* differences. The findings of this chapter show that comprehension varies dependent on a variety of human variables, and any change in the usual wording should be carefully tailored to the target readers. The limitations of the experiment and some future work which would lead to improvements were discussed in Section 5.5.2. Section 5.5.2 also presented the original idea of C-factor as a unified factor for evaluating reading comprehension.

Next, Chapter 6 will present the evaluation of the impact of the CLCM simplification on tasks important for the domain.

Chapter 6 – The impact of the CLCM Simplification on other tasks

Without translators, Europe would not exist; translators are more important than members of the European Parliament. (Milan Kundera)

Say what we may of the inadequacy of translation, yet the work is and will always be one of the weightiest and worthiest undertakings in the general concerns of the world. (J. W. Goethe)

The chapter will present the second approach to evaluation, namely investigating the impact of CLCM on extrinsic tasks. The "Translation and Post-editing Experiment" will be described. This experiment was run on emergency instructions for the General Population, employing a publicly available Machine Translation (MT) engine and twenty-five volunteer translation specialists. The chapter is composed as follows. Section 6.1 will present the general motivations for this investigation. Section 6.2 will introduce the related work on evaluating the impact of CL on translation tasks. Section 6.3 will describe the setting of the experiment. Section 6.4 will provide the

results, Section 6.5 will discuss the findings, and finally Section 6.6 will present the conclusions.

6.1. Introduction, Definitions and Motivations

Previously, Chapter 5 showed that the use of the controlled language CLCM (described in Chapter 4), which has been developed specifically for the Crisis management document-type *Instructions for the General Population* (discussed in Chapters 3 and 4), has a positive impact on text simplification as measured by metrics, such as diminishing Text Complexity and increasing Text Comprehensibility of documents re-written according to the CLCM guidelines.

However, in order to fully assess the impact of CLCM, it is necessary to test its effect on other extrinsic tasks. This chapter provides an extrinsic evaluation of the controlled simplification approach by measuring its impact on tasks which are important for the domain. These tasks are manual translation (ManT) and machine translation (MT).

This thesis defines translation as the process of "transferring a written text from source language to target language, conducted by a translator, in a specific socio-cultural context" (Hatim, 2004). A distinction is made between Manual Translation (ManT), Computer-Aided Translation (CAT), and (fully automatic) Machine Translation (MT). MT is defined as "the use of computers to automate some or all of the process of translating from one language to another" (Jurafsky and Martin, 2008). The CAT tools' aim is to assist human translators in their work without replacing them completely (Bowker, 2002). Typical CAT tools are the Translation Memory systems (TM) (Bowker, 2002), which work by splitting the source text into segments and storing their translation correspondents for future re-use.

The translation tasks are important for the CM domain because in the modern global world, emergency instructions need to be translated to other languages as they need to be used in other

countries besides the country (or language) of origin, or in order to improve non-native speakers' access to them. This can especially be seen in websites with world-wide impact, such as www.redcross.org[39]. For this reason, evaluating the impact of CLCM on manual and machine translation is considered to be important. Additionally, previous studies have shown that CLs can cut translation costs between 50 to 70% (Pym, 1993).

The decision to test the impact of CLCM on fully automatic or computer-assisted machine translation systems, in addition to the evaluation on ManT, has been dictated by the fact that NLP applications are gaining more and more ground in translation (Kittredge, 2003) because they increase its speed. In fact, studies have shown that MT followed by post-editing is 40% faster than manual translation (Sousa et al., 2011).

The choice for testing the CLCM impact on an MT system rather than on a TM or another CAT system has several motivations. One of them is the fact that although CAT tools are widely used by translation specialists nowadays, (at least partial) MT is gradually replacing them. In fact, there are recent studies showing that human translators post-editing MT had a significantly higher productivity and the translation quality was higher than the work of the same translators editing TM fuzzy matches (Guerberof, 2009). Another motivation is that CAT tools are usually available only to translation specialists and not the the general public. MT engines (e.g. Google Translate) are both available to the general public, who are the target readers of the *Instructions for the General Population,* and are widely used for the translation of websites.

Next, Section 6.2 will present the related work in evaluating manual and machine translation.

39 Last accessed on December 11th, 2010.

6.2. Related Work in Evaluating CL on Manual and Machine Translation

This section will introduce the related work on evaluating Controlled Languages (CL) on external tasks employing translation. In translation studies, the original text that needs to be translated is called *source text*, while the text obtained as an end-product of the translation – *target text* (Al-Qiani, 2000).

To the knowledge of the author, there are no published approaches evaluating the impact of controlled languages on Manual Translation. This can be explained by the fact that the existing CLs aim at improvement of human comprehension (Section 2.3.2.1), Machine Translation (Section 2.3.2.2), and at both human comprehension and MT (Section 2.3.2.3), but not of the manual translation.

There are a limited number of approaches evaluating the impact of CL on MT and some of them employ the existing MT evaluation methods. The common element between all approaches is that comparison between the complex and the simplified text is usually conducted. The approaches can be divided into those employing post-editing (PE) and those which do not employ post-editing of the MT output.

The post-editing approaches can again be classified according to the taxonomy of approaches of Krings (2001), which divides the approaches into temporal, technical, and cognitive evaluation methods. Only a few of them involve a combination of quality evaluation of the MT engine, such as BLEU scores (Papineni et al., 2002; Aikawa et al., 2007), and PE evaluation metrics. Also, only a

few of them conduct evaluation from all three perspectives. Krings (2001) measures the post-editing effort from temporal (time necessary to post-edit a text), technical (number of additions, deletions, and cuts-and-pastes, i.e. edit distance) and cognitive points of view (the think-aloud protocols introduced above). O'Brien (2005, 2006) also measures the time and number of additions, deletions, and cuts-and-pastes, but employs Choice Network Analysis for measuring the cognitive effort. In contrast, Aikawa et al. (2007) evaluated the PE only from a technical point of view (character-based edit distance) according to the afore-mentioned classification and mainly relies on human evaluation scores and BLEU scores (Papineni et al, 2002). All of the afore-mentioned studies conclude that CL pre-editing can improve the quality of machine translation output. A study that does not employ any post-editing is Vassiliou et al. (2003), who has conducted a modification-focussed comparison of the errors produced in the MT outputs for both complex and simplified texts.

The difference between the previous evaluation approaches and the work presented in this thesis is that the CL evaluation approaches presented in this section evaluate CL for technical documentation, while the evaluation presented here evaluates a CL for documents employed in crisis management communication (for a description of crisis management communication, see Section 1.1). Due to the motivations laid down in Section 6.1, which make machine translation preferable to manual translation, the evaluation of CLCM on machine translation will be much wider than on manual translation. More specifically, the evaluation perspective taken to measure the impact of CLCM on machine translation will follow the perspective of Krings (2001). The evaluation on manual translation will be accordingly based on temporal evaluation. The evaluation methods chosen to evaluate the CLCM impact on manual and machine translation, together with the settings of the experiment, will be provided in Section 6.3.

6.3. Settings of the Translation and Post-editing Experiment

In order to evaluate the impact of CLCM on ManT and MT, an experiment (the *"Translation and Post-editing Experiment"*) has been conducted. The aim of the experiment was to evaluate the impact of CLCM on ManT by comparing the time employed for manually translating the complex and the simplified texts, and on MT by comparing the post-editing cost for the complex and the simplified texts. The PE evaluation has been conducted from all three existing perspectives (Krings, 2001): temporal, technical, and cognitive points of view. The materials used in the experiment included two texts, one of which was simplified according to the CLCM guidelines; twenty-five translation specialists, used both for manual translation and for MT output post-editing; a specially developed web interface; and an MT engine. This section will provide all of the details regarding the way the experiment was set up. Section 6.3.1 will present the research hypotheses which were investigated, Section 6.3.2 will describe the texts used, Section 6.3.3 will present details about the simplification rules employed for producing the simplified text, Section 6.3.4 will discuss the texts' preparation, Section 6.3.5 will provide details about the translation specialists who took part in the experiment, and Section 6.3.6 will provide details about the MT engine used for the MT evaluation, as well as details regarding the post-editing and manual translation instructions provided to the participants. Finally, Section 6.3.7 will present the web interface developed for the experiment.

This experiment was conducted in collaboration with Dr. Constantin Orasan. His contributions were the edit-distance calculation for the technical evaluation of the post-edited MT output (described in Section 6.4.2.2) and the web interface used for the experiment (described in Section 6.3.6).

6.3.1. Research hypotheses investigated

The aim of the experiment was to test the following research hypotheses:

1. **CLCM has a positive impact on manual translation.**

2. **CLCM has a positive impact on machine translation.**

The first hypothesis was tested by measuring the time employed by human translators to manually translate the simplified and the complex texts and by comparing their results. The assumption was that if CLCM had a positive impact on manual translation, then the time that translation specialists employ for translating the simplified text will be less than the time spent manually translating the complex text.

The second hypothesis was tested by automatically translating both the simplified and the complex texts with an MT engine and comparing the machine translations of both texts from three evaluation perspectives: time employed to post-edit the text, edit distance between the MT output and the post-edited text, and cognitive effort involved in post-editing the MT output text. The testing of the second hypothesis is based on the following three assumptions:

1. If CLCM has a positive impact on the MT engine performance, then the average time that human post-editors spend on correcting the MT output of the simplified text will be lower than the average time spent post-editing the MT output of the complex text.

2. If CLCM has a positive impact on the MT engine performance, then the average edit distance

between the MT output text and the post-edited text will be lower for the simplified text than for the complex text.

3. If CLCM has a positive impact on the MT engine performance, then the cognitive effort required for the human post-editors to post-edit the simplified text will be less than the cognitive effort required for post-editing the complex text.

6.3.2. Description of the texts used

Two texts were used for the experiment. For clarity, the two original texts will be referred to as *Text 1-original* and *Text 2-original*, while the simplified forms of the texts will be referred to as *Text 1-simplified* and *Text 2-simplified*. The texts were extracted from the same source document, "Individual Preparedness and Response to Chemical, Radiological, Nuclear, and Biological Terrorist Attacks" (Davis et al., 2003). The source document is a large (thirty-five pages) guide edited by the RAND Corporation[40]. It contains an analysis of strategies adopted and suggestions for individual guidelines for actions to be taken during four types of terrorist attacks: chemical, radiological, nuclear, and biological. The instructions are defined by the authors as "defined in terms of simple rules that should be easy for individuals to adopt". The document is considered as appropriate for the experiment for the following reasons:

- Terrorist attacks are currently considered to be a very sensitive topic.
- The document is addressed to the general public.
- As terrorist attacks often involve a large mass of people of different nationalities, a translation of the document would most probably be necessary.

40 http://www.rand.org/, last accessed on January 13th, 2011.

- The individual strategies suggested have also been included in the document as leaflets to be spread world-wide, thus resulting in four leaflets: for chemical, radiological, nuclear, and biological attacks. Both texts for the experiment were taken from these leaflets. Both leaflets have been attached to the present thesis in Appendix D. More concretely, *Text 1-original* was taken from the leaflet entitled "Nuclear Attack". It deals with strategies for avoiding radioactive fallout. *Text 2-original* was taken from the leaflet "Chemical Attack", which contains instructions for how to find clean air very quickly. Both leaflets have the same structure. They first outline the goal to be reached and then list the specific actions to be undertaken in order to reach the goal while avoiding difficult situations.

The texts were purposely selected so as to be of a similar length. The length of the original texts as taken from the documents are:

- Text 1-original (Nuclear Attack): 140 words
- Text 2-original (Chemical Attack): 138 words

The two texts were selected so as to be of a similar length in order to ensure balanced comparability when conducting the experiment. The texts' length was kept down to around one hundred fifty words for each text in order to not overload the participants, who were volunteers, while still ensuring sufficient text for testing the research hypotheses. The texts were selected in such a way that they exhibit a similar quantity of issues considered to be problematic both for MT engines (long sentences, anaphora) and human translators (e.g. specialist terminology). Even though Steedman (2008) has demonstrated some of the weaknesses of the MT engine used (Google Translate, see Section 6.3.6.) using a very small informal experiment, such as long-range dependences (like object

relative clauses and long-distance anaphora), as well as local-range errors (such as lexical ambiguity), little is known about the complete range of concrete TC issues, as there are no publications regarding the evaluation of Google Translate nor regarding the types of errors it produces. Given this and since Google changes without any notice the MT algorithm itself, and as it was difficult to directly apply the MT engine-specific factors that negatively affect Mtranslatability according to Bernth and Gdaniec (2001), only general knowledge about what constitutes a TC issue for NLP applications was used (See Table 2.1 in Section 2.1.3). The two texts were analysed for TC issues by applying the TC analysis presented in Chapter 3.

Similarity in Text Complexity has thus been guaranteed, to the extent possible, by two main factors:

1. Provenance from the same source document (same genre, same style of language, and same author)
2. Similar or same length

Tables 6.2 and 6.3 provide the results of the TC analysis run on the two texts. The texts were analysed only for presence of Main and Secondary TC features. Since splitting of long, complex sentences into shorter and simpler ones is considered to be one of the main and most basic TS operations, the number of sentences has been also considered a TC criterion, with lower number of sentences being a marker of higher TC.

TC issue\Text	**Text 1-original (Nuclear Attack) Original**	**Text 2-original (Chemical Attack) Original**
Number of sentences	13	16
Average sentence length (in words)	10.769	8.625
Average word length (in letters)	4.807	4.913

Lexical diversity (types/tokens)	0.493	0.608
Average number of word senses	7.868	9.631
Proportion of function words	0.378	0.297

Table 6.2: Comparison of the Main TC features analyses of *Text 1-original* and *Text 2-original.*

TC issue\Text	**Text 1-original (Nuclear Attack) Original**	**Text 2-original (Chemical Attack) Original**
Proportion of Coordination markers/word-tokens ratio	0.036	0.058
Proportion of Subordination markers/word-tokens ratio	0.086	0.058
Proportion of Relative markers/word-tokens ratio	0.007	0.0
Proportion of Ambiguous quantifiers/word-tokens ratio	0.014	0.0
proportion of punctuation signs/ all tokens ratio	0.140	0.168
Proportion of Discourse markers/word-tokens ratio	0.007	0.014
Proportion of Pronouns/word-tokens ratio	0.014	0.022

Table 6.3: Comparison of the Secondary TC features analyses of *Text 1-original* and *Text 2-original.*

As can be seen, both Tables 2 and 3 are structured in the same way. The first column lists the Main TC features (in the case of Table 6.2) and the Secondary TC features (in the case of Table 6.3), while the second column in both tables lists the results of each of these TC features for *Text 1-original* and the third column lists the same for *Text 2-original*. As can be seen in Table 6.2, although both texts have a similar number of sentences and similar average sentence length, *Text 2-original* has a higher number of shorter sentences than *Text 1-original*. It can be also seen that *Text 1-original* is characterized by a lower lexical diversity index and a lower number of senses per word than *Text 2-original*. On the other hand, *Text 2-original* exhibits a slightly lower proportion of function words. Table 6.3 also shows differences between the original versions of *Text 1-original* and *Text 2-original*. They are:

- *Text 2-original* exhibits a higher proportion of coordinate markers and a lower proportion of subordinate markers than *Text 1-original*.
- Both texts have a very low proportion of relative markers and ambiguous quantifiers, with *Text 2-original* having none of either of them.
- While both texts have a similar number of punctuation signs, *Text 2-original* has a higher number of discourse markers and personal and possessive pronouns than *Text 1-original*.

For the purposes of the experiment, *Text 1-original* was left as it is in the source document, while *Text 2-original* was manually simplified according to the principles of CLCM. Next, Section 6.3.3 provides details of the method of simplification applied to *Text 2-original* in order to obtain *Text 2-simplified*.

6.3.3. Method of simplification

As was stated earlier, for the purposes of the experiment, *Text 2-original* (Chemical Attack) was simplified according to the CLCM rules, obtaining the result of the simplification, *Text 2-simplified*.

The simplification was performed manually by the author of the present thesis, following the principles of a prototype version of CLCM. The simplification did not follow any translation target language-specific rules. The prototype version of CLCM contained a subset of the rules of the current version of CLCM. More concretely, the rules used for simplification were the following:

- Use only literal meaning.

- Avoid idiomatic expressions.
- Use concrete (instead of abstract) concepts.
- Write short sentences.
- Write only one piece of information (condition, instruction, or item) per line.
- Use the allowed structure ‘How to …’ for writing titles.
- Divide the specific situations into separate blocks.
- Write a title for every specific situation.
- Remove unimportant information.
- If an adjective modifies more than one entity, repeat the adjective next to every modified entity.
- Write conditions before the corresponding instructions.
- Use less ambiguous expressions.
- Avoid technical terms.
- If possible, use a finite verb instead of an ‘-ing’ form.

As can be seen, a difference between the prototype version of CLCM and the current one was the rule “Remove unimportant information”, which was removed in the current version due to the wish to preserve the information content of the original text. The prototype version of CLCM was created after a dry-run with undergraduate students in order to estimate what kind of rules would positively affect the performance of the selected MT engine.

A comparative TC analysis for *Text 2-original* and *Text 2-simplified* is provided in Table 6.4. As a criterion for low TC, the number of sentences has also been provided. The comparison between the text lengths of *Text 2-original* and *Text 2-simplified* is given below:

- Length of *Text 2-original*: Number of words: 138
- Length of *Text 2-simplified*: Number of words: 134

As can be seen, the simplification did not reduce much and did not extend at all the length in words of the original text. In order for the TC analysis to be run, the two texts were pre-processed like the whole corpus, using the Connexor parser and then the Python scripts described in Chapter 3.

TC issue\Text	**Text 2-original**	**Text 2-simplified**
Main TC features		
Number of sentences	16	20
Average sentence length	8.625	6.7
Average word length	4.913	4.179
Lexical diversity	0.608	0.470
Average number of word senses	9.631	12.562
Proportion of function words	0.297	0.388
Secondary TC features		
Proportion of coordination markers/word-tokens ratio	0.058	0.015
Proportion of subordination markers/word-tokens ratio	0.058	0.045
Proportion of relative markers/word-tokens ratio	0.0	0.007
Proportion of ambiguous quantifiers/word-tokens ratio	0.0	0.0
Proportion of punctuation signs/ all tokens ratio	0.168	0.152
Proportion of discourse markers/word-tokens ratio	0.014	0.0
Proportion of pronouns/word-tokens ratio	0.022	0.089

Table 6.4: Comparison of all of the TC features of *Text 2-original* and *Text 2-simplified*.

As can be seen, Table 6.4 is composed in a similar way to Table 6.2 and Table 6.3. The first column contains the TC features which have been examined, while the second and the third column contain respectively the results for *Text 2-original* and *Text 2-simplified*. As can be seen from the table, there are situations of decreased TC, no changes in TC, and increased TC. A discussion of these issues follows below:

1. Positive impact:

   - The number of sentences in *Text 2-simplified* has increased, while the number of words has remained the same, which means that the average length of the sentences has decreased (also shown in the next row), which is a clear indication of diminished TC (Klebanov *et al.,* 2004).

   - The average word length has also decreased, which can be explained by the replacement of technical terms by more common ones, as well as lexical diversity, which means that there is an increased consistency of the employed terminology. All of these indicate decreased TC of *Text 2-simplified*.

   - The decrease in lexical diversity, together with the increase in the proportion of function words (meaning a decrease in the proportion of content words, thus leading to a decrease in vocabulary richness), are both indicators of decreased TC.

   - The proportion of coordination markers, the proportion of subordination markers, and the proportion of punctuation signs to the total number of tokens in the text have also decreased, which is an indication of decreased sentence length and sentence complexity,

and thus also an indication of decreased TC.

2. No changes:

   - The proportion of ambiguous quantifiers stayed the same, as no ambiguous quantifiers were present in *Text 2-original* and none were added to *Text 2-simplified*.

3. Negative impact:

   - The average number of word senses per word has increased, which can be explained by the use of more common terms being more ambiguous (Zipf, 1949).

   - The proportion of relative markers has increased, which may lead to an increased TC, and may also be explained as being an effect of insufficiently concrete simplification rules (e.g. “write short sentences” instead of “split relative clauses into a separate sentence”).

   - The proportion of discourse markers to the word-tokens has decreased, but this can be explained by the fact that there was no rule suggesting the use of more discourse connectives in the prototype CLCM version.

   - Finally, the proportion of personal and possessive pronouns has increased, which was again due to the lack of a rule regarding their omission.

Next, Section 6.3.4. will provide further details about the preparation of the texts for the

experiment.

6.3.4. Preparation of the texts

The texts which were finally used for the experiment were *Text 1-original* and *Text 2-simplified.* They were used in their plain text format. In order to be prepared for the experiment, the two texts were split into alternating sentences. Some sentences were left in English, while others were translated into the seven target languages using the chosen MT engine. Figure 6.1 shows a screenshot of the interface, in which the sentence splitting with alternating English and Spanish sentences are given from *Text 1-original*, while the sentence splitting and the obtained MT translations per language are given in Appendix D.

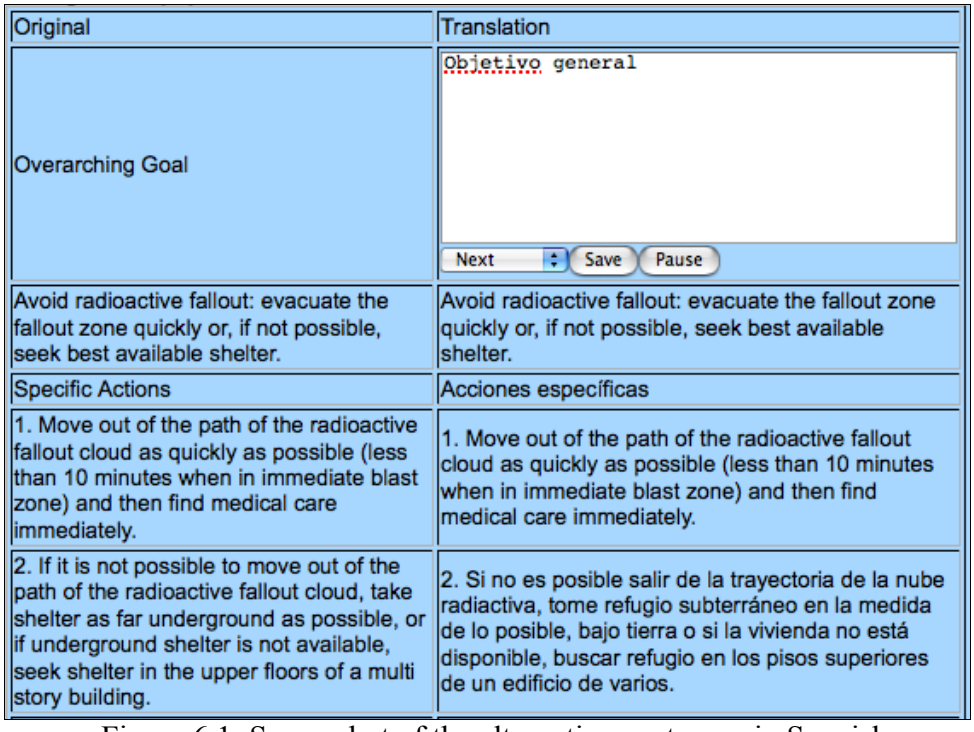


Figure 6.1: Screenshot of the alternating sentences in Spanish.

The left column of Figure 6.1 shows the English *Text 1-original* split into alternating sentences, while the right column shows the alternating sentences, with the first one in Spanish, the second one

in English, the third one in Spanish again, etc. The reason for alternating between original and translated sentences was to prevent translators from getting used to the terminology as they translate a text. The drawback of this alternation was that the humans involved in the experiment had to alternate between two different activities (translating and post-editing), which may have slowed down their reactions. Next, Section 6.3.5 will provide details about the participants in the experiment.

6.3.5. Participants

Twenty-five translation specialists with at least four years of professional translation experience were involved in the experiment. The translation specialists were either freelancers working with translation agencies, or translators from the European Parliament Directorate General for Translation in Luxembourg[41].

The translators involved in the experiment were experts in translating from English to their native languages, which were in total seven Indo-European languages written in three writing systems (Cyrillic, Latin and Greek alphabets):

- Three Slavic languages: *Bulgarian*, *Slovenian*, and *Russian*
- One Romance language: *Spanish*
- one Germanic language: *Dutch*
- One Semitic Language: *Maltese*
- Modern Greek

41 http://www.europarl.europa.eu/parliament/expert/staticDisplay.do?id=54&pageRank=9&language=EN, last accessed on January 15th, 2011.

The distribution of participants per language involved a minimum of three participants and a maximum of five participants per language. The different number of participants varied according to how many specialists agreed to participate in the experiment. The ages of the participants were between thirty and fifty-six years old. Section 6.3.6 will provide details about the MT engine used, as well as the instructions given to the participants in the matters of translation and post-editing.

6.3.6. Machine translation engine, post-editing and translation instructions

The MT engine selected for the experiment was Google Translate[42], a freely available online statistical MT engine developed by Google Inc. Currently, Google Translate offers translations between fifty-eight languages. It is widely used to translate short texts or websites.

In order to obtain the MT translations, the alternating sentences from the source English *Text 1-original* and from the source English *Text 2-simplified*, which were aimed to be post-edited, were manually copy-pasted into the Google Translate online interface (shown in Figure 6.2), then translated automatically into the seven target languages described in Section 6.3.5, and finally copy-pasted into the appropriate fields of the online web interface, described in Section 6.3.7.

42 http://translate.google.com/, last accessed November 18th, 2010.

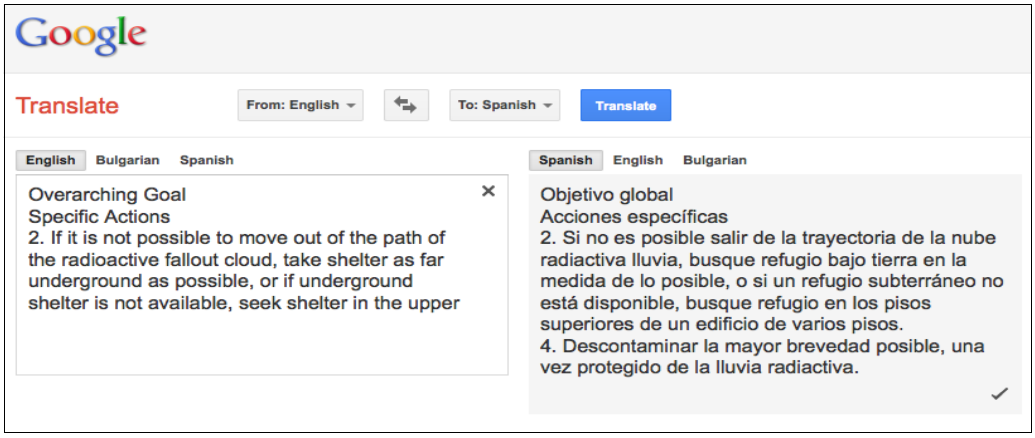


Figure 6.2: Screenshot of the online interface of the MT engine Google Translate.

Figure 6.2 shows the Google Translate translation of the alternating sentences to be post-edited, taken from *Text 1-original* translated from English (on the left) to Spanish (on the right).

After obtaining the MT translations of the appropriate sentences, and before starting the experiment, the participants were provided with instructions regarding the method to follow for translating the sentences left in English and the method for post-editing the sentences automatically translated into their target language. Regarding manual translations, the instructions provided to the participants were limited to information regarding the purpose of the final translated documents and their readers, namely that the translations were not aimed for publication and that the target readers were from the general population and usually not specialists in the domain. For this reason, the participants were instructed to keep the style close to everyday language and not to do lengthy searches for the correct technical term, but rather to use a general term instead.

More attention was paid to post-editing, since the participants involved in the experiment had sound experience in manually translating text documents but no experience in post-editing MT output texts, except for experience in proofreading manually translated texts produced by other translators. In fact, according to Schäfner (2003), there is a difference between post-editing and proofreading,

with proofreading being the last step of the post-editing process. The instructions regarding the method of post-editing which were provided to the participants were based on the existing types of post-editing applied by post-editors to the MT output text. The guidelines given to the post-editors in the present experiment followed suggestions for full post-editing (consisting in complete re-writing of the MT output into naturally sounding texts), as well as the instructions given in Wagner (1985) for the use of European Commission post-editors.

In the first place, as for the manual translation task, the post-editors were given instructions regarding the purpose of the post-editing (i.e., that the edited document will not be published). Also, post-editors were given the instruction to ignore stylistic errors if they didn't affect sentence meaning. Finally, they were instructed to write in a clear and easily understandable style and also to avoid using idiomatic expressions in the post-edited text.

The instructions given to the participants in this experiment are provided in Appendix 3 of the present thesis. Finally, Section 6.3.7 will provide details about the interface used in the experiment.

6.3.7. Interface used

The experiment employed a specially designed web interface, developed by Dr. Constantin Orasan.

The interface allowed entering user-specific data in the first step (Figure 6.3), such as user name, and then choosing between pairs of source and target languages in the second step, and the two texts in the third step (Figures 4a and 4b). Then, a screen displaying the sentences to translate and post-edit (as already shown in Figure 6.1) was displayed. In particular, as can be seen in Figure 6.5, after editing or translating each of the sentences, three options are given: to go to the next sentence, to go

to the previous sentence, or to finish the whole text. There were also a button to save the changes and a button to pause the timer in case an interruption is needed. Finally, the participant is prompted to enter his/her level of English and the level of the respective target language and was allowed to see user statistics (Figure 6.6).

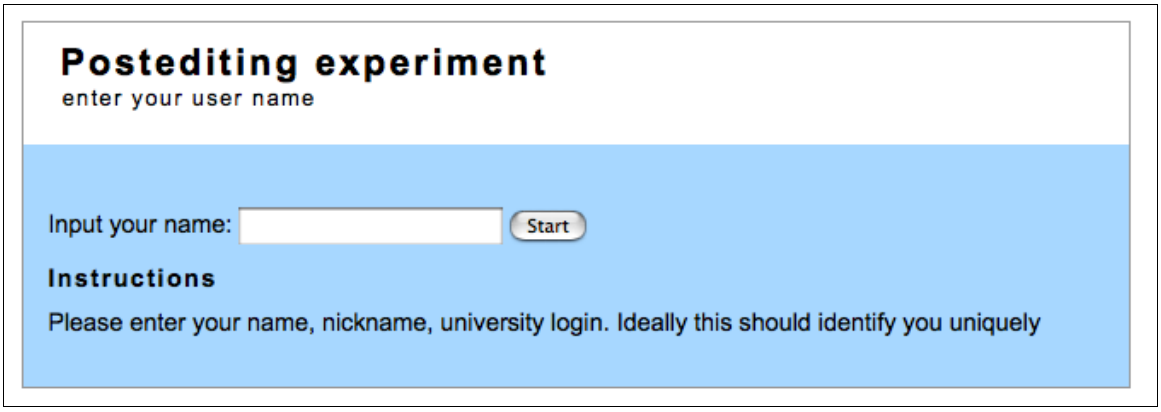


Figure 6.3: Entering user-specific data.

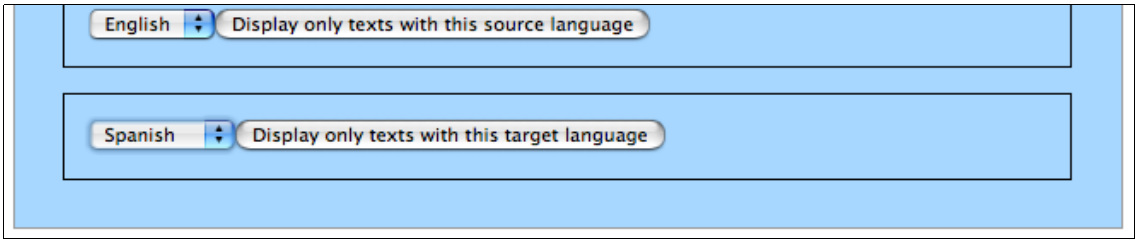


Figure 6.4a: Choosing a language pair.

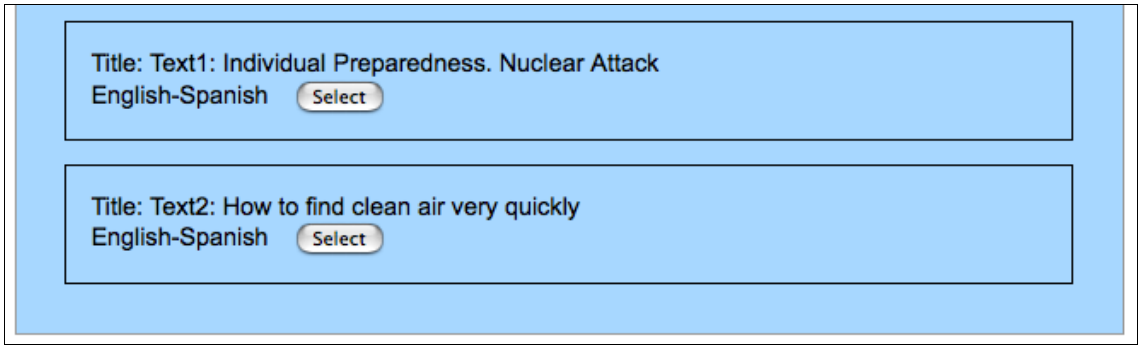


Figure 6.4b: Choosing a text.

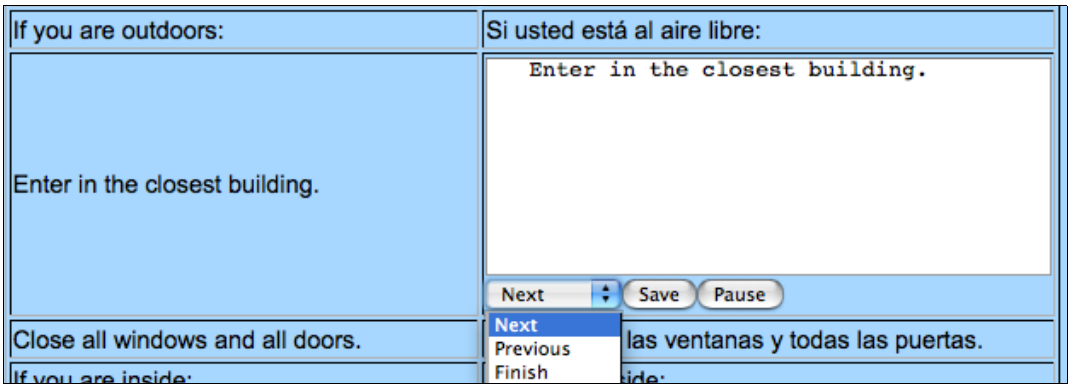


Figure 6.5: Moving between sentences.

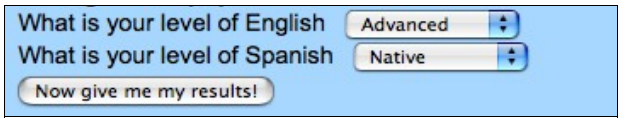


Figure 6.6: Obtaining statistics.

The web interface was tested with the help of final year undergraduate students of Linguistics and Computational Linguistics at the University of Wolverhampton, UK. Next, Section 6.4 will present the results obtained from the experiment.

6.4. Evaluation Methods and Results

As stated in Section 6.3.1, the evaluation of the impact of CLCM on translation tasks consisted of testing the following research hypotheses:

1. CLCM has a positive impact on manual translation.

2. CLCM has a positive impact on machine translation.

As has also been explained, the first hypothesis will be tested by comparing the results of measuring the time employed by human translators to manually translate the simplified and the complex texts,

while the second hypothesis will be tested by comparing the post-editing of the machine translations of both texts from three different evaluation perspectives: time employed to post-edit the text, amount of changes between the MT output and the post-edited text, and cognitive effort involved in post-editing the MT output text.

This section will present the results obtained in this experiment. More concretely, Section 6.4.1 will present the results obtained from the evaluation of the effect of CLCM on ManT, while Section 6.4.2 and its sub-sections will presents the results relative to the effect of CLCM on MT.

The evaluation of the impact of the CLCM simplification on ManT in Section 6.4.1. and the two first methods of evaluation of the impact of the CLCM simplification on MT in Sections 6.4.2.1. and 6.4.2.2. are based on a peer-reviewed published article (see Temnikova and Orasan, 2009) but a more precise evaluation has been made and the discussion has been significantly expanded.

6.4.1. Evaluation method and results of the impact of CLCM on manual translation

The method applied to evaluate the impact of CLCM simplification on manual translation actually represented a method of evaluating whether the simplification had a positive impact on the process of manual translation, and thus relates to the ease of translation of the source text (also defined "source text translatability" and not to the quality of the translation which was obtained.

The evaluation approach which was applied was an empirical one—measuring the amount of time necessary for the human translators to manually translate the sentences left in English in *Text 1–original* and in *Text 2-simplified* and comparing their results. The time was measured in seconds.

The statistical significance of the difference of means has been calculated. As stated in Section 6.3.2, the assumption was that if CLCM had a positive impact on manual translation, then the time that translation specialists employ for manually translating the simplified text will be less than the time spent manually translating the complex text.

The evaluation results are listed in Table 6. The values presented are the mean ± standard error of the mean. The means are the averages per language of the time employed by the different translators to translate each sentence into the respective target language. As the parts of the two texts (*Text 1-original* and *Text 2-simplified*) which the translators had to translate manually had different lengths (characteristics of original and simplified texts and their length are given in Table 5), the time employed was normalised by the length of the English sentence (calculated in number of characters). In this way the values in Table 6 represent time employed in seconds divided by sentence length in characters, i.e. measured in seconds/character. Table 5 presents the alternative sentences, chosen to be left for manual translation from English to the seven target languages.

Text/Details	**Text**	**Size**
Text 1– original	Avoid radioactive fallout: evacuate the fallout zone quickly or, if not possible, seek best available shelter. 1. Move out of the path of the radioactive fallout cloud as quickly as possible (less than 10 minutes when in immediate blast zone) and then find medical care immediately. 3. Find ways to cover skin, nose, and mouth, if it does not impede either evacuating the fallout zone or taking shelter. 5. If outside the radioactive fallout area, still take shelter to avoid any residual radiation.	Words 83 Characters 496
Text 2 – simplified	Attack outdoors Enter in the closest building. If you are inside: Go upstairs. Seal the room. When you are told that it is safe to go out: Go out immediately. Follow the chemical attack plans for your building. Open the windows.	Words 63 Characters 338

	If you can not open the windows: OR When you are protected from the chemical agent: Take a shower. Seek medical treatment.	

Table 5: Texts and text length of the two texts for ManT.

As can be seen, Table 5 is composed of three columns and three rows. The rows indicate the respective text, while the second column contains the selected sentences left for the translators translate manually and the third column provides numerical details about the size of the text in words and characters. As can be seen, there are substantial differences between the sizes of the selected sentences of *Text 1-original* and *Text 2-simplified* (496 vs. 338 characters), which motivates the necessity of normalizing the time per length of sentences in characters. The division per number of characters instead of dividing per number of words is motivated by the wish for consistency with the normalisation of edit distance values, described in Section 6.4.2.2.

Target language/Text	**Text 1-original (sec/char)**	**Text 2-simplified (sec/char)**	**Difference**
Bulgarian	0.977 ± 0.134	0.879 ± 0.123	+ 0.098
Dutch	1.287 ± 0.087	1.390 ± 0.089	- 0.103
Russian	1.397 ± 0.127	1.407 ± 0.099	- 0.010
Greek	1.397 ± 0.235	1.426 ± 0.101	- 0.029
Slovenian	1.007 ± 0.298	0.694 ± 0.085	+ 0.313
Spanish	1.708 ± 0.249	1.319 ± 0.298	+ 0.389
Maltese	1.001 ± 0.109	0.741 ± 0.059	+ 0.260
All	1.234 ± 0.075	1.125 ± 0.076	+ 0.109

Table 6.6: Time employed to manually translate texts.

As can be seen, Table 6.6 is composed of four columns and nine rows. The first column contains the target languages into which the translators translated the sentences left in English in each of the two texts. The second through eighth rows contain data for the individual languages, while the last row contains the data for all of the languages. The second column contains the data for *Text 1-original* and the third column contains the data for *Text 2-simplified*, while the fourth column contains the

difference between each value in the second column and the respective value in the third column, with positive numbers showing a positive impact and negative numbers a negative impact of the CLCM simplification on the manual translation of *Text 2-simplified*. The values for all languages for the manual translation task as a whole are given as an indication of the overall impact of the CLCM simplification on this task, despite the fact that the languages are very different. The values have been rounded to the third digit after the decimal point. As already stated in Section 6.3.5, the number of translators per language varied between three and five, with the number of translators for Bulgarian and Dutch being five, while the number of translators for the rest of the languages was three, with twenty-five translators for all languages in total.

As can be seen in Table 6.6, second column, the highest value is the time to translate *Text 1-original* from English is to Spanish, while the lowest value is to translate the same text to Bulgarian. The values for Russian and Greek are the same, when rounded to the third decimal place. On the other hand, as can be seen in the third column, the highest value is the time to translate *Text 2-simplified* from English to Greek, while the lowest value is for translation of the same text to Slovenian. The averaged time per character to translate *Text 1-original* thus ranges between 0.977 and 1.708, while the averaged time per character to translate *Text 2-simplified* ranges from 0.694 to 1.426. As can be seen, the range of values for the simplified text is lower than the range of values for the original text, which supports the first hypothesis. It can also be seen that although three of the languages (Dutch, Russian and Greek) have higher values for the simplified text, the rest of the languages, as well as the values for all of the languages as a whole, have substantially lower values for the simplified text, which again supports the first hypothesis. The statistical hypothesis testing of the differences of means of the original-simplified pairs of times have shown weak evidence for the obtained results, with the maximum for the means of all of the languages as whole of 84% confidence and with the exception of the means of Maltese with 97% confidence using 1-tailed,

directional t-test. It has been assumed that the reason for such low statistical significance are the very small sample sizes and it has been calculated that, in order to obtain stronger evidence, larger samples are necessary. Namely, in order to obtain 95% confidence it is necessary to have at least sixty-four translators for all of the languages as a whole and between nine and forty-seven translators for the single languages.

A discussion of these results will be presented in Section 6.5.1. Next, Section 6.4.2 will present the results obtained for the impact of the CLCM simplification on machine translation.

6.4.2. Evaluation method and results of the impact of CLCM on machine translation

The evaluation presented in this section aims to test hypothesis N. 2: that CLCM has a positive impact on machine translation.

The approaches applied to evaluate the impact of the CLCM simplification on the process of MT constitute methods of evaluation of the quality of the obtained translation, and more concretely of measuring the cost involved in manually post-editing MT output text.

The impact of CLCM on the cost of post-editing has been evaluated from three perspectives: the temporal point of view, the technical point of view, and the cognitive effort point of view. Section 6.4.2.1 will introduce the method of temporal post-editing evaluation and its results, Section 6.4.2.2 will present the method of technical post-editing evaluation and its results, and finally Section 6.4.2.3 will present the method of evaluating the cognitive effort involved in post-editing the MT output text and its results for a subset of the original target languages.

6.4.2.1. Temporal evaluation of post-editing

The evaluation of the impact of the CLCM simplification on the cost of full post-editing involved in re-writing of the MT output in naturally-sounding language from the temporal point of view consists of comparing the time employed by human post-editors in re-writing the sentences of *Text 1-original* and *Text 2-simplified* obtained as Google Translate output. As stated in Section 6.3.2, the assumption related to the evaluation of the impact of CLCM on post-editing from the temporal perspective was that if CLCM has a positive impact on the MT engine performance, then the average time that human post-editors spend on correcting the MT output of the simplified text will be lower than the average time spent post-editing the MT output of the complex text.

Similarly to the evaluation of manual translation, the time was measured in seconds and then normalised per number of characters of each translated sentence. However, in this case there is a difference in the normalisation, in that the time recorded has been divided by the characters of the MT output in the respective language of the sentences selected for MT rather than by the characters of the English input text. The results are shown in Table 6.8, while Table 6.7, similarly to Table 5, shows the sentences of *Text 1-original* and *Text 2-simplified* selected for translation by Google Translate, as well as the lengths of the MT output in the seven languages. The MT outputs of the MT inputs in all of the target languages, as has already been pointed out in Section 6.3.4, are provided in Appendix D of this thesis.

Text/Details	**Text in English**	**Size of translation**	
		Language	**Size**
Text 1 – original	Overarching Goal Specific Actions 2. If it is not possible to move out of the path of the radioactive fallout cloud, take shelter as far	**Bulgarian**	Words 52 Characters 347
		Dutch	Words 53 Characters 362
		Russian	Words 45

	underground as possible, or if underground shelter is not available, seek shelter in the upper floors of a multi story building. 4. Decontaminate as soon as possible, once protected from the fallout.		Characters 304
		Greek	Words 55 Characters 369
		Slovenian	Words 47 Characters 302
		Spanish	Words 59 Characters 346
		Maltese	Words 45 Characters 340
Text 2 – simplified	How to find clean air very quickly If you are outdoors: Close all windows and all doors. Stay inside. Find an interior room. Stay inside until you are told it is safe to go out. Ventilate the room. Attack inside If there are no chemical plans for your building: Breathe fresh air. Go out to the street. Go to the roof. Remove your clothes. When it is safe to go out:	**Bulgarian**	Words 62 Characters 381
		Dutch	Words 72 Characters 386
		Russian	Words 60 Characters 386
		Greek	Words 70 Characters 415
		Slovenian	Words 63 Characters 363
		Spanish	Words 70 Characters 410
		Maltese	Words 56 Characters 366

Table 6.7: Texts in English and the lengths of their MT translations.

As can be seen, Table 6.7 is composed of four columns and three rows, the rows indicating the respective text, while the columns contain the information relative to the respective text. More concretely, the second column contains the input text in English, while the second and the third column contain the MT output language and the size of the MT output text. The size of the text is again provided in number of characters; additionally, the number of words is given here. As can be seen, the lengths of the different output translations of the same input text vary substantially (between 302 and 369 characters for *Text 1-original* and between 363 and 415 characters for *Text 2-simplified*), which motivates the necessity for normalisation of the time employed to post-edit by the length of the MT output texts. The time is normalised per number of characters instead of per number of words, as post-editing can also occur at the character level, and thus the number of characters is more important in this evaluation than the number of words. The averaged normalised times to post-edit each of these MT output texts are provided in Table 6.8.

Target language/Text	Text 1-original (sec/char)	*Text 2-simplified (sec/char)*	Difference
Bulgarian	0.865 ± 0.118	0.624 ± 0.134	+ 0.241
Dutch	1.088 ± 0.155	0.948 ± 0.117	+ 0.140
Russian	1.649 ± 0.312	1.135 ± 0.103	+ 0.514
Greek	0.863 ± 0.058	0.607 ± 0.087	+ 0.256
Slovenian	1.002 ± 0.359	0.600 ± 0.068	+ 0.402
Spanish	1.014 ± 0.124	0.476 ± 0.042	+ 0.538
Maltese	0.734 ± 0.186	0.669 ± 0.280	+ 0.065
All	1.022 ± 0.082	0.733 ± 0.062	+ 0.289

Table 6.8: Time employed to manually post-edit MT output texts.

Similarly to Table 6.6, Table 6.8 is composed of four columns and nine rows. The first column contains the target languages into which some of the sentences of the two texts were automatically translated. More concretely, the second through eighth rows contain the individual languages, while the last row contains the data for all of the languages. The second column contains the data for *Text 1-original* and the third column contains the data for *Text 2-simplified.* Finally, the fourth column contains the difference between each value in the second column and the respective value in the third column, with positive numbers showing a positive impact and thus a decrease of the time employed to post-edit the simplified text in comparison with the original text, while negative numbers show a negative impact, and thus an increase of the time to post-edit the simplified text compared to the original text. The values for all languages for the task of post-editing the machine translation output as a whole are given as an indication of the overall impact of the CLCM simplification on this task, despite the fact that the languages are very different and thus the errors generated by the MT engine should be very different too. The values have been rounded to the third digit after the decimal point. As the post-editors and translators are the same people, the number of post-editors per language and in total for all languages is the same, and namely for Bulgarian and Dutch the post-editors are five while for the rest of the languages – three, with twenty-five post-editors in total for all the seven languages.

As can be seen from Table 6.8, in contrast with Table 6.6, there is a clear decrease in the time to post-edit the simplified text for all output languages, which strongly supports the second hypothesis.

More concretely, as can be seen in Table 6.8, second column, the highest value to post-edit *Text 1-original* is for the Russian MT output, while the lowest time is the time to post-edit the Maltese MT output translation, with values thus ranging from 0.734 sec/char to 1.649 sec/char. The range of values in the third column, for *Text 2-simplified,* is between 1.135 and 0.476, with the highest normalised time to post-edit the simplified text again being that for Russian, while the lowest time is the one for Spanish. The high value for Russian may be explained by the type of language and that in fewer context more errors are created and the morphological level.

The smallest differences are for Maltese and Dutch, which means that the impact of CLCM simplification on post-editing MT-translated text is smaller for these two languages.

The difference between the normalised means for *Text 1-original* and *Text 2-simplified* for all of the post-edited sentences from English to all of the target languages is significant at 99% confidence limits using the 1-tailed (directional) t-test. Except for Greek (95% confidence) and Spanish (99% confidence), as in the case of the time employed to manually translate the two texts, the statistical hypothesis testing of the differences of means of for the individual languages shows weak evidence that CLCM has a positive impact on manually post-editing texts. In order to obtain more precise statistical results for the separate languages, an experiment with bigger samples (currently the entries are three or five per language) is necessary. It has been calculated that in order to reach a statistical significance of at least 95%, the number of post-editors need to be more than five. Discussion of the results will be presented in Section 6.5.2. Next, Section 6.4.2.2 will present the results obtained for the impact of the CLCM simplification on machine translation from the

technical point of view.

6.4.2.2. Technical evaluation of post-editing

As has been stated before, the impact of the CLCM simplification on MT has been tested from three perspectives: temporal, technical, and cognitive. This section aims to describe the evaluation of the CLCM impact on MT from the technical point of view. The technical evaluation also aims to test the second research hypothesis, namely that CLCM has a positive impact on machine translation.

The technical evaluation of post-editing consists of measuring the amount of changes necessary for the post-editor to post-edit the MT output text in order to transform it into naturally-sounding language. As stated in Section 6.3.2, the assumption related to the evaluation of the impact of CLCM on post-editing from the technical point of view is that if CLCM has a positive impact on MT engine performance, then the average edit distance between the MT output text and the post-edited text will be lower for the simplified text than for the complex text. In order to calculate the amount of changes necessary, the Levenshtein edit distance (Levenshtein, 1966) between the MT outputs and the post-edited versions was calculated for both texts. The edit distance calculation was performed by Dr. Constantin Orasan, while the author of the thesis performed the rest of the calculations (i.e. statistical significance). The Levenshtein edit distance is a metric used in computer science that provides a numerical value of the difference between two strings by calculating the minimal number of edits (insertions of one character, deletions of one character and replacements of one character) necessary to transform one string into another. The results of this evaluation are provided in Table 6.9 and represent the means of the edit distance numbers ± the standard error of the mean. The edit distance values have been normalised by dividing them by the number of characters in each edited sentence, similarly to the normalisation of the times to translate and post-

edit described in Sections 6.4.1 and 6.4.2.1. Thus, the values are listed in edits per character (edit/char). The means are again rounded to the third digit after the decimal point, in order to facilitate comparison between the tables of results of the different evaluation techniques.

Target language/Text	**Text 1-original (edit/char)**	***Text 2-simplified (edit/char)***	**Difference**
Bulgarian	0.328 ± 0.051	0.225 ± 0.030	+ 0.103
Dutch	0.467 ± 0.076	0.440 ± 0.050	+ 0.027
Russian	0.365 ± 0.087	0.324 ± 0.043	+ 0.041
Greek	0.329 ± 0.092	0.181 ± 0.031	+ 0.148
Slovenian	0.388 ± 0.080	0.380 ± 0.068	+ 0.008
Spanish	0.158 ± 0.068	0.198 ± 0.036	- 0.040
Maltese	0.758 ± 0.116	0.393 ± 0.060	+ 0.365
All	0.390 ± 0.030	0.310 ± 0.020	+ 0.080

Table 6.9: Normalised edit distance values between the MT output and the post-edited texts.

As can be seen, similarly to Tables 6 and 8, Table 6.9 is composed out of four columns and nine rows. The first column contains the target languages into which the translators translated the sentences left in English in each of the two texts. The second through eighth rows contain the data for the individual languages, while the last row contains the data for all of the languages, which are provided in other to show overall performance. The second and third columns contain, respectively, the data for *Text 1-original* and *Text 2-simplified*. Finally, the fourth column contains the difference between each value in the second column and the corresponding value in the third column, with positive numbers showing a positive impact and negative numbers showing a negative impact of the CLCM simplification on post-editing the MT translated simplified text. As can be seen, eight out of the nine differences in the fourth column are positive, which supports the second hypothesis and indicates that the CLCM simplification had a positive impact on most of the language pairs, except for the post-editing of the English-Spanish Google Translate output. The second column shows that the range of values for *Text 1-original* is 0.158–0.758 edit/char, with the highest value being for Maltese and the lowest value being for Spanish. This means that more edits of the MT-translated

original text were needed in Maltese than in Spanish. The third column shows that the range of values for *Text 2-simplified* is smaller and that it is between 0.181 edit/char and 0.440 edit/char, with the highest value being for Dutch and the lowest value being for Greek. This means that the Dutch MT-translated simplified text needed the most post-editing, while the MT-translated Greek simplified text needed substantially less post-editing. The fact that the range of values for the simplified text is smaller than the range of values for the original text shows that there is less variation and thus supports the second hypothesis—that the CLCM simplification has a positive impact on manually post-editing MT output text. There is a large variation in differences between the first and the second column means, as can be seen in the fourth column. This means that the impact of CLCM strongly depends on the pair of languages, and can be explained by the different qualities of the Google Translate engine pairs. As can be seen, the least positive impact that CLCM has is on Slovenian, while the largest positive impact of CLCM is on post-editing the Maltese MT output data.

Statistical significance testing shows very high confidence levels for the difference of means for all languages as a whole (98.7%), Bulgarian (95.7%), Maltese (99.7%), and Greek (97.4%); medium confidence levels for Russian (67.7%), Spanish (69.8%), and Dutch (61.0%); and very low confidence levels for Slovenian (52.3%). It is hypothesized that the low confidence levels are due to the fact that the samples are very small, and it has been calculated with larger samples (a minimum of nine post-editors) the confidence will increase. The high statistical significance results also support the second research hypothesis—that the CLCM simplification has a positive impact on MT post-editing.

Next, Section 6.4.2.3 will present the evaluation of the impact of the CLCM simplification on MT from the cognitive point of view.

6.4.2.3. Evaluation of the cognitive effort involved in post-editing

As has been seen in Sections 6.4.2.1 and 6.4.2.2, the temporal and technical MT evaluation approaches provided positive, but inconclusive results. More concretely:

1. Although time shows a decisive improvement in post-editing of the simplified text, edit distance shows some conflicting results (e.g. a negative result for Spanish not existing in the temporal results).

2. Different combinations of languages benefit highly from simplification from the two evaluation perspectives (Spanish and Russian benefit the most for time; Maltese and Greek benefit the most for edit distance).

3. Different combinations of languages benefit less from simplification from the two evaluation perspectives (Dutch and Maltese for time; Dutch and Slovenian for edit distance).

For these reasons, and as the need for following a triangulation method in evaluation has been seen, it was considered that there is need for a third MT evaluation perspective which would throw light on the concrete challenges of post-editing the machine translations of the original and the simplified texts. Triangulation (Denzin, 1970) is a method used in the social sciences to collect data or evaluate collected data in a more holistic way. It is motivated by the fact that one evaluation method may mislead, two evaluation methods may clash, and adding a third evaluation method may help to see the whole picture of the data better, as well as regarding the different evaluation results not as separate but as one whole picture of the elements interacting between them.

The third and final evaluation approach again aims to test the second hypothesis, namely that CLCM has a positive impact on machine translation. The innovative evaluation methodology presented below has been peer-reviewed with very positive reviews and published (Temnikova, 2010).

The evaluation methodology is based on the assumption that if CLCM has a positive impact on MT engine performance, then the cognitive effort required for the human post-editors to post-edit the translation of the simplified text will be less than the cognitive effort required for post-editing the translation of the complex text.

In order to choose the best method for evaluation of cognitive effort, the limitations of the existing approaches have been studied. There are three existing evaluation methods for measuring the cognitive effort involved in post-editing: the think-aloud protocols (TAP), the choice-network analysis (CNA) and the manual rating of translations according to the level of difficulty involved in post-editing them (Sousa, et al., 2011). The limitations of the existing cognitive effort evaluation approaches which motivated the development of a new method were:

- As TAP consists of post-editors commenting on their decisions out loud (Krings, 2001), its shortcoming is that it cannot easily be formalised and reused.

- As CNA (O'Brien, 2005; O'Brien, 2006) focuses on the number of different changes that the post-editors apply to MT output words, considering that the larger the number of different changes of the same word, the more cognitive effort a post-editor needs to undertake in order to choose an option, the shortcoming of this evaluation technique lies in the fact that it is not certain that all options are equally available to all post-editors.

- The limitation of manually ranking translations (Sousa, et al., 2011) according to their difficulty of post-editing is that it is very subjective and does not shed light on the concrete difficulties that the post-editors encounter while correcting MT output text.

Due to the shortcomings of the three existing cognitive evaluation methods, a new evaluation approach is proposed here. As the new approach is a revolutionary one, the evaluation was initially carried out only on a subset of the previous testing data, i.e. on only three languages: Bulgarian, Russian, and Spanish. The choice of languages which entered into the testing subset was motivated both by the need to test the evaluation approach on different language types and by the availability of evaluators. The languages chosen were from two different language families. Bulgarian is considered to be distinctive in the fact that it exhibits characteristics of both a Slavic and a Balkan language; Russian is known to be a highly inflected Slavic language, while Spanish is an analytic language from the Romance family.

The new method consisted of examining the MT output texts for types of errors in terms of the difficulty of correcting them and comparing the number of occurrences of the different classes of errors in the translations of *Text 1-original* and *Text 2-simplified*. In relation to the assumption stated earlier, it is considered that if the MT output of *Text 1-original* exhibits more errors which are cognitively difficult to correct than the MT output of *Text 2-simplified,* this will support the hypothesis that CLCM simplification has a positive influence on MT. The development of the new evaluation approach is based on three steps:

1. First, an MT error classification was adopted.

2. Second, the adopted MT error classification was enriched with information regarding the

cognitive difficulty of manually correcting the different types of errors.

3. Third, an experiment involving human evaluators annotating MT generated errors and counting such errors was conducted.

The original contribution of this evaluation approach is that this is the first evaluation approach measuring the cognitive efforts of post-editing MT output, employing an objective pre-defined MT error correction taxonomy which is adaptable to other texts. Previous approaches have either been pre-defined error classifications based on a general typology of MT-generated errors, without any weighting of how difficult these errors are to correct (Schäfer, 2003; Allen, 2003), or lists of types of post-editing operations based on the empirical analysis of the input and output of post-editing the concrete texts (Tatsumi, 2010). The novelty which the new approach contributes to the knowledge and resources in this domain is that it joins together a general MT error classification with error weights, based on the cognitive efforts that post-editors experience, which makes this evaluation approach both capable of providing an objective numerical measure of the post-editing process and straightforwardly applicable to other texts and domains.

The MT error classification adopted for this experiment was a modified version of Vilar et al.'s (2006) MT error classification, which classifies MT generated errors into four main categories: missing words (1), word order (2), incorrect words (3), and punctuation errors (4). Some of the main categories are further divided into a few sub-categories. The classification is shown in Table 6.10.

Error	**Correction**
(1.) Missing word	Correcting the error requires supplying the missing word.
(2.1.) Word order error	Correcting the error requires changing the order of single words.

(2.2.) Word order error	Correcting the error requires changing the order of whole phrases.
(3.1.) An incorrect word	Correcting the error requires replacing a word with a completely different lexical item.
(3.2.) Correct word with an incorrect ending (e.g. number or case)	Correcting the error requires replacing the ending of a word with a different ending.
(3.3.) An incorrect word	Correcting the error requires replacing the word with a stylistically different synonym.
(3.4.) Extra word	Correcting the error requires removing the extra word.
(3.5.) Error due to incorrectly recognised idiomatic expressions	Correcting the error requires replacing the word with the correct translation of the idiomatic expression.
(4.1.) Missing punctuation sign	Correcting the error requires adding the missing punctuation sign(s).
(4.2.) Incorrect punctuation sign	Correcting the error requires replacing the wrong punctuation sign(s) with the correct punctuation sign(s).

Table 6.10: MT error classification.

As can be seen, Table 6.10 is composed of two columns, the first one containing the type of error and the second one describing the operations required from post-editors in order to correct it.

The next step after adapting the existing MT classification was that the original MT output error types were further divided into classes on the basis of the cognitive effort involved in manually detecting and correcting these kinds of errors. The error correction classes were based on the cognitive model of reading (Harley, 2008) already described in Section 2.1.1; Baddeley's working memory theory (Baddeley & Hitch, 1974); and written error detection studies (Larigauderie, 1998). Working memory plays an important role in reading and thus also in correcting written errors. According to (Harley, 2008), working memory is composed of a central executive (which plays the role of a supervisory system), a phonological loop, and a visuo-spatial temporal store. Written language comprehension is performed unconsciously in several steps based on collection of information from the previous stages of processing (Baddeley & Hitch, 1974; Harley, 2008), with different combinations of working memory components involved at different stages. The first processing stage to be performed is grapheme recognition, followed by the process of lexical access, and finally by the process of syntactic and semantic processing. The first two processing stages are considered to be less cognitively costly because they only require activation in memory

of previous representations and mental vocabulary look-up, while syntactic and semantic processing on the one hand make an additional brain component active (the phonological loop); and on the other hand, experiments (Larigauderie, 1998) have shown that the syntactic and semantic levels challenge the central executive much more, as they involve understanding the whole text passage and relating its meaning to the meanings deduced from the previous text passages. Reading difficulty also depends on the processing sentence span, generating in this way the following ranking of sentence segments in order of increasing reading difficulty: word level, clause level, and sentence level. The process of correcting written errors is strictly related to the process of reading and also requires processing the information from the different stages of reading. This evaluation approach is based on the assumption that the post-editing task is very similar to an appositely performed error detection and correction task. According to this point of view, the less cognitively costly errors should be those at the word level, i.e. words with incorrect endings, which require only a grammar rule representation look-up, and the most cognitively expensive ones should be those involving syntactic and semantic processing of the whole sentence.

On the basis of these theories of the cognitive effort of reading and correcting written errors, a relative ranking of MT errors, according to the relative ease of correction, was constructed. The types of errors are divided into three groups, according to the cognitive errors correction theory—namely into the morphological level, the lexical level and the syntactic level. Each of the levels contains from one to four types of errors. Table 6.11 shows the Vilar et al. (2006) MT error classification, enriched with the cognitive effort ranking, with (1) being the easiest error to correct and (10) the hardest error to correct.

Level	**Type of error**
Morphological level	1. Correct word, incorrect form (CInF)
	2. Stylistically incorrect synonym (Styl)

Lexical level	3. Incorrect word (InW)
	4. Extra word (ExW)
	5. Missing word (MissW)
	6. Idiomatic expression (Idiom)
Syntactic level	7. Wrong Punctuation (InP)
	8. Missing Punctuation (MissP)
	9. Word Order at Word level (WoW)
	10. Word Order at Phrase level (WoPh)

Table 6.11: Cognitive effort MT error ranking.

As can be seen Table 6.11, similarly to Table 6.10, is composed of two columns. The first column contains the different linguistic levels of errors while the second column contains the list of errors, ordered from the easiest to correct (Correct word, incorrect form) to the most difficult to correct (Word Order at Phrase level). Next to each of the categories is provided an abbreviation, which will be used further on due to space limitations. For example, *Missing word* is abbreviated as *MissW*. As can be seen, the errors are listed in order starting from those requiring local corrections at the character or morphological level (such as re-writing a correct word, in the correct form); and ending up with changes at phrase/clause and sentence level (for example moving a word from one position to another, for example in the Russian "Когда это безопасно выходить на улицу:" /*When it is safe to go out*/, re-written as "Когда выходить на улицу безопасно:" /*When going out is safe*/).

An experiment which supports this error ranking is the recent contribution of Tatsumi (2010), who, after examining the concrete process of post-editing technical documentation, has discovered that correcting punctuation is "one of the most effort intensive" post-editing tasks. In fact, according to the post-editing error ranking proposed in this thesis, correcting punctuation is one of the types of MT error correction that require one of the highest levels of cognitive effort, due to the fact that before correcting punctuation the whole or at least part of the sentence must be processed and its meaning understood.

After the error classification was adapted and ranked according to the cognitive effort involved in correcting each kind of error, an experiment involving human evaluators was conducted. The experiment was based on the analysis of the post-edited versions of the two texts of a random post-editor.

The evaluation involved three independent evaluators (one per language) who were native speakers of the respective languages. The evaluators were asked to manually analyse the post-edited versions of the MT outputs of the two texts, classify and annotate the different types of errors, and provide additional explanations. The evaluators were given guidelines (provided in Appendix D together with the results for the three languages), which specified the steps to follow and the necessary information to provide.

The results of the number of errors for each of the three languages of the subset are provided in Figures 7, 8, and 9 respectively for the Spanish, Bulgarian, and Russian languages. Inter-annotator agreement was not calculated as there was only one annotator per language and text version.

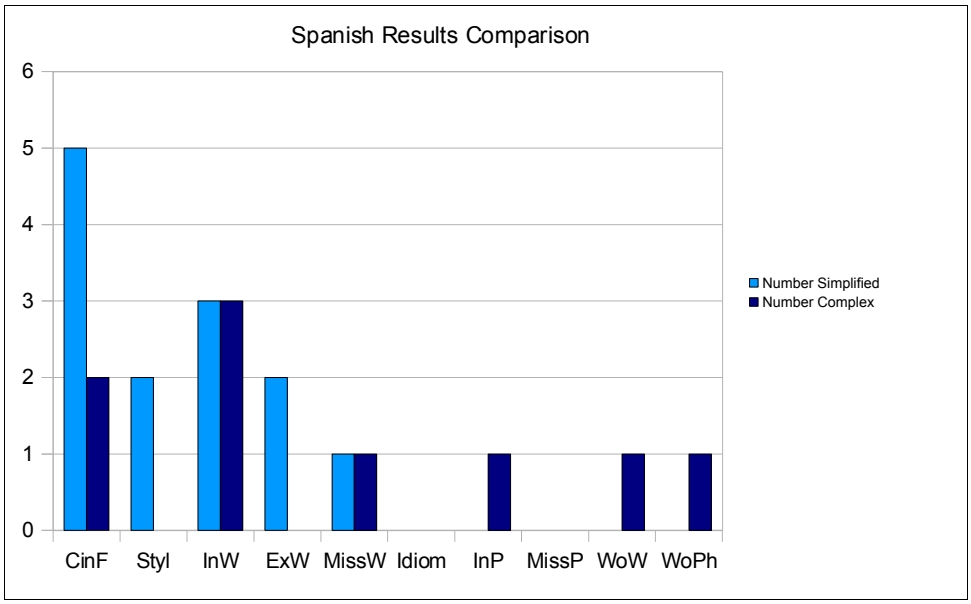


Figure 6.7: Post-editing error distribution for Spanish text.

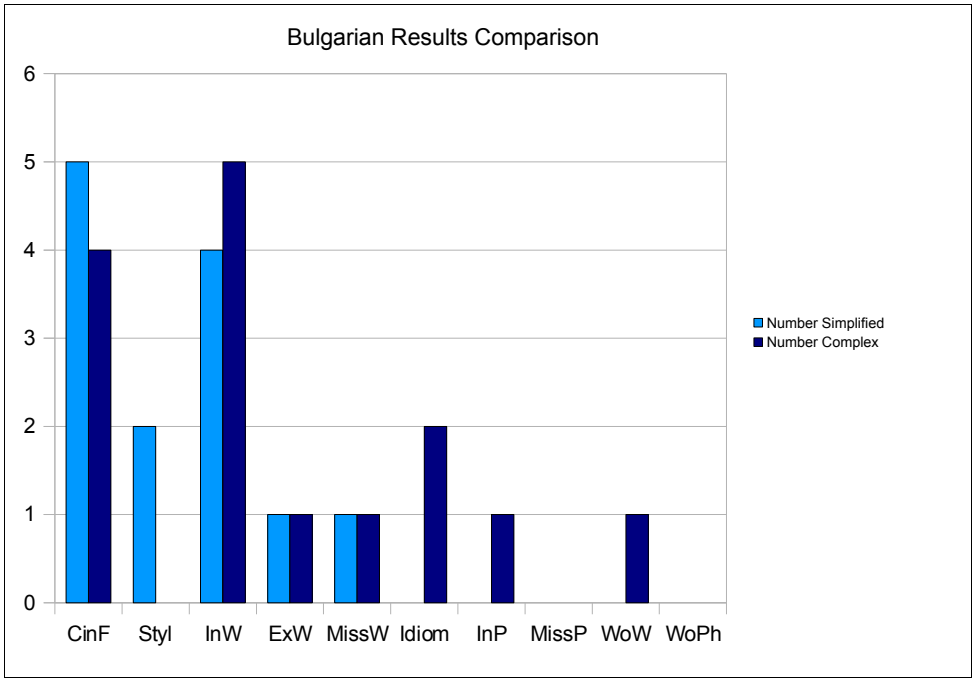


Figure 6.8: Post-editing error distribution for Bulgarian text.

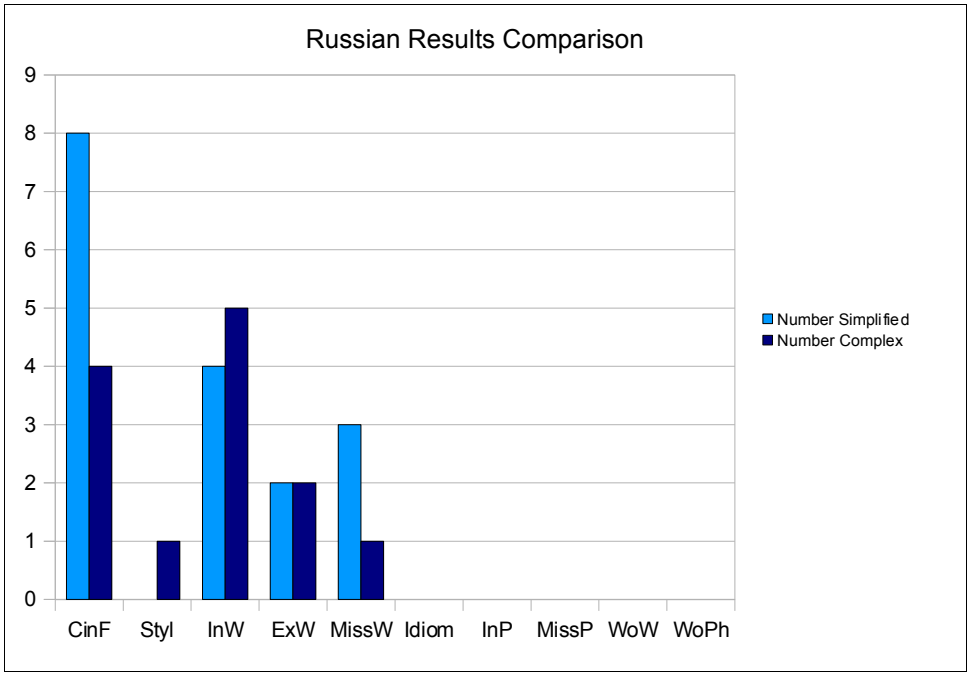


Figure 6.9: Post-editing error distribution for Russian text.

As can be seen, the three graphics are composed in the same way. The MT output post-editing errors are listed on the X axis, starting from the errors considered to be the simplest ones to correct from the cognitive effort point of view (Correct Word Incorrect Form) on the left, continuing with the lexical ones in the middle (Incorrect Style Synonym, Incorrect Word, Extra Word, Missing Word

and incorrect Idiomatic Expression), and ending with the hardest ones to correct from the cognitive point of view of the post-editor involved, i.e. Wrong Punctuation, Missing Punctuation, Word Order at Word Level and Word Order at Phrase Level. The number of errors is given on the Y axis; they vary between zero and five for the Spanish and Bulgarian results and between zero and eight for the Russian results. The coloured columns represent the number of errors of each type, with the dark columns indicating the number of errors in the translation of *Text 1-original* and the bright columns indicating the number of errors in the translation of *Text 2-simplified*.

As can be seen, the distribution of errors for Spanish and Bulgarian are similar, probably due to the fact that both languages are analytic. A difference is noted in the results for Russian, which is a synthetic language (i.e. relying more on endings). In fact, as can be noticed in Figure 6.9, there are no errors in Russian in the second half of the graphic from left to right. This means that the evaluator considered that there were no errors at the syntactic level in either text. The lack of word-order-related MT errors in Russian in both texts can be explained by the fact that since Russian is considered to be a relatively free-word-order language, unusual word order is not considered to be a mistake by either the evaluators or the post-editors, as it can be explained by stylistic variations.

A comparison between the number of errors of *Text 1-original* and *Text 2-simplified* shows that despite the fact that there are errors of all types in *Text 1-original* (dark columns), the distributions of errors in *Text 2-simplified* (bright columns), are situated in the left half of the graphics, despite the fact that the bright columns seem higher than the dark ones. This clear difference in error distributions means that there are no syntactic-type errors in the translation of *Text 2-simplified*, in comparison with the translation of *Text 1-original,* which shows all types of errors. This finding shows that *Text 2-simplified* has errors which are easier to correct, which supports research hypothesis N.2—that CLCM has a positive impact on MT. The fact that the bright columns in the

left part of the graphics are higher means that there is a higher number of easier-to-correct MT errors (at the morphological level and most easiest at the lexical level). This can be explained by the fact that CLCM simplification results in very short segments with very little context, which hinder the MT engine disambiguation of the ending of the word or the word meaning. As has been mentioned, however, the Russian language is synthetic and thus ambiguity of morphological endings should play a more important role for it than for the other two languages. An example of this problem is the fact that instructions in controlled language do not normally contain a subject. For example, "Stay inside." is translated wrongly into Russian as "Пребывание внутри." ("Staying inside"). Adding the subject improves the translation. ("You should stay inside." is translated correctly).

It is considered that due to the small sample size, further study with a larger sample is necessary in order to properly evaluate the impact in the case of the Russian language and whether any modification of the evaluation theory is appropriate for synthetic languages.

The fact that there are no stylistic errors in the complex texts can be explained by the fact that post-editors were explicitly given the instruction to avoid considering stylistic errors if there is no drastic change in the meaning. The increased number of 'missing and extra words' in the Russian and Spanish simplified texts can be explained by the fact that less context generates more syntactical ambiguity and thus more syntactic errors. There are no 'idiomatic expression' errors in the Spanish and Russian cases, while errors of this kind are present in the Bulgarian complex text and disappear in the Bulgarian simplified text. Their disappearance can be explained by the instruction given to post-editors to specifically avoid using idiomatic expressions.

Due to the low number of errors and the small numerical difference between them in the two texts,

no statistical significance of population means or of difference of means was calculated. Next, Section 6.5 will summarize the research findings of this experiment.

6.5. Summary of the Findings and Discussion of the Results

This section will present the discussion of the results presented in Section 6.4 and the findings of the experiment. Section 6.5.1 will present the findings regarding the evaluation of the impact of CLCM on Manual Translation, while Section 6.5.2 will present the findings obtained while testing the impact of CLCM on Machine Translation. Section 6.5.3 will outline the differences, and finally, Section 6.5.4 will list some of the external factors which may have influenced the analysis.

6.5.1. Impact on manual translation

As stated in Section 6.3.1, the first research hypothesis investigated in this experiment was:

1. CLCM has a positive impact on manual translation.

The experiment to test this hypothesis was based on the assumption that if CLCM has a positive impact on manual translation, then the time that translation specialists employ for translating the simplified text will be less than the time spent manually translating the complex text.

The results presented in Section 6.4.1, Table 6.6 showed that although three out of the seven languages showed an increase in the time for the simplified text to be manually translated, compared to the time for the original (unsimplified) text to be manually translated, the following

observations were also true:

- The results for the average of all of the languages showed a decrease in translation time for the simplified text.

- The results for the rest of the languages, with the exception of the three that showed an increase, showed a decrease in translation time for the simplified text.

- When a decrease was observed, it was relatively low (0.029-0.103) compared to the increase for the rest of the languages (0.098-0.389).

Although the statistical analysis of the differences between the means does not show a high level of confidence, this was attributed to the small sample size and to the fact that translators were penalised by not being able to employ any dictionaries or other resources; thus, it can be considered that the first hypothesis is supported by the experiment and that CLCM has a positive impact on manual translation.

6.5.2. Impact on machine translation

According to Section 6.3.1, the second research hypothesis investigated in this chapter was that:

2. CLCM has a positive impact on machine translation.

The testing of the second hypothesis is based on the following three assumptions:

1. If CLCM has a positive impact on MT engine performance, then the average time that human post-editors spend on correcting the MT output of the simplified text will be lower than the average time spent post-editing the MT output of the complex text.

2. If CLCM has a positive impact on MT engine performance, then the average edit distance between the MT output text and the post-edited text will be lower for the simplified text than for the complex text.

3. If CLCM has a positive impact on MT engine performance, then the cognitive effort required for the human post-editors to post-edit the simplified text will be less than the cognitive effort required for post-editing the complex text.

The results presented in Sections 6.4.2.1, 6.4.2.2, and 6.4.2.3 have shown that:

- All normalised mean times show a decrease in post-editing time for the simplified text, with confidence levels of 99% for the mean of all of the languages as a whole, which confirms the second hypothesis.

- The single language pairs do not show an acceptable statistical significance of the difference of means, which is considered to be caused by the very low sample sizes.

- The normalised edit distance values show a decrease in the edit distance for post-editing the simplified text for the mean of all of the languages as a whole with confidence levels of 98.7%, which supports the second hypothesis.

- Some of the single languages pairs do not show an acceptable statistical significance of the difference of means, which is considered to be caused by the very low sample sizes.

- The comparative analysis of the MT error distributions, conducted because of the need for a third evaluation measure, showed that there is a clear decrease in the number of cognitively difficult-to-correct errors in the simplified text compared to the non-simplified text, which like the findings of the temporal and technical evaluations, supports the second hypothesis.

All of the evidence presented above supports the second hypothesis—that the CLCM simplification has a positive impact on post-editing machine translation.

6.5.3. Comparison between the findings related to manual and machine translation

Comparing the results of the impact of CLCM on manual and machine translation, the following findings have emerged:

- Comparing the time results for manual translation, where for three out of the seven languages CLCM has a negative impact, with all of the results for MT, it seems that MT is more positively influenced by CLCM than ManT.

- The comparative analysis of the times employed to manually translate and to post-edit the MT outputs of the two texts in Table 6.6 and Table 6.8, confirm the results of Sousa et al. (2011), that the MT translation + post-editing is faster than manual translation for both complex and simplified texts. More concretely, the average time normalised per number of

characters employed by all participants to post-edit the complex text was 17.2% less than the average normalised time to translate the complex text. This difference increased to 34.8% for the simplified text. Both differences of the means were statistically significant with 98-99% confidence.

- An additional analysis has shown that the standard deviations for ManT are the same for the two texts, while for MT the standard deviations are as a whole smaller for the simplified text. This can mean that all of the post-editors in all languages are experiencing better effects of CLCM simplification than is the case when translating the texts.

These findings lead to the conclusions that MT is better influenced by the CLCM simplification than ManT, which is a good discovery because currently the use of MT is increasing and also because it has been proved that MT+ post-editing is faster than manual translation.

6.5.4. Influence of external factors

On the basis of the experiments run and further analysis of the results, which showed a large variety of findings, it was concluded that some factors may have influenced the results, including the following:

- Translators employed no additional resources, e.g. dictionaries or terminological databases, to assist them with translation. A further study with more appropriate simulation of translators work is desirable.

- The human post-editing factor: professional translators do not like to post-edit automatically

translated texts, especially when the quality of translation is not sufficiently high.

- Interface shortcomings: the users did not like the fact that they had to alternate between original texts and automatically translated sentences. In fact, they have also suggested improvements in order to make the interface more user-friendly.

- The experiment was not conducted in a controlled environment, so some of the post-editors and translators could have been distracted by external factors while performing the task.

- No evaluation of the Google Translate language pairs has been conducted. The observed variability may also be due to this factor.

An attempt will be made to try to restrict all of these factors and to explore their influence in future work. Next, Section 6.6 will present the conclusions of this chapter.

6.6. Conclusions

The aim of the present chapter was to test the impact of the CLCM simplification method presented in Chapter 4 on external tasks. The external tasks selected were Manual and Machine translation, as translation of emergency instructions is important in the modern global world. MT has been selected instead of translation memories because of the fact that MT is available to both translation specialists and others and is known to be gradually replacing TM (Guerberof, 2009). The evaluation was based on two research hypotheses, namely that CLCM has a positive impact on manual translation (research hypothesis 1) and also has a positive impact on machine translation (research hypothesis 2). The hypotheses have been tested via the *"Translation and Post-editing Experiment"*

involving twenty-five human translators and post-editors, a pair of original and simplified texts, an online MT engine, and a specially designed translating and post-editing interface.

The impact of CLCM on ManT has been measured on the basis of the time translators employ to translate the two texts, while the impact on MT was measured via three different evaluation methods, all aiming to evaluate the post-editing cost, but in different ways. The three different ways were the time employed to post-edit the two texts, the edit distances between the MT output and the post-edited texts, and the amount and types of errors which are easy or difficult to manually correct in both MT output texts.

The results, laid out in Sections 6.4.1 and 6.4.2, have shown that CLCM has a positive impact on both manual and machine translation, with better results seen for machine translation from all three evaluation perspectives. In addition, the results confirmed the already published discovery that post-editing MT is faster than manually translating text with larger difference for the simplified text. A set of external factors influencing the experiment have been identified and discussed in Section 6.5.4. Future work will include testing the evaluation approach on larger data sets, more languages, a statistical significance analysis and with more restrictions of the external factors.

Next, Chapter 7 will present the evaluation of the internal process of manual text simplification according to the CLCM guidelines, as well as a study of which operations may need to be automated to assist human simplifiers in their work.

Chapter 7 – Qualitative and Quantitative Analysis of the CLCM Simplification Process

Make things as simple as possible, but not simpler. (Albert Einstein)

The chapter describes the last experimental evaluation of CLCM, referred to as the "Text Simplification Task Experiment". This experiment investigates the acceptability to users of the CLCM guidelines and sheds light on the cost of the manual simplification process. The structure of the chapter is as follows. Section 7.1 provides the motivations for this experiment. Section 7.2 will present the general setting of the experiment. Section 7.3 will introduce the methodology and the results of the evaluation relative to estimating the cost of manual simplification. Section 7.4 will describe the methodology for and results of evaluating the difficulty of applying concrete CLCM rules, and Section 7.5 will investigate the priorities for future implementation. On the basis of the previous sections, Section 7.6 will give a summary of the findings of this experiment. Finally, Section 7.7 will present the conclusions and plans for future work.

7.1. Introduction

This chapter presents the last of the three evaluation perspectives of the Controlled Language for Crisis Management (CLCM), presented in Chapter 4. As was stated in Section 2.3.3, previous findings show that manual simplification according to controlled language (CL) rules is a very cognitive-effort-intensive and time-consuming process (Goyvaerts, 1996; Huijsen, 1998) and series of CL editors have already been created (Renahy et al., 2010; Schwitter et al., 2003). Nothing has been reported regarding the concrete difficulties encountered during manual simplification nor about any investigation of the user requirements prior to implementing CL aiding tools. This motivates the evaluation described in this Chapter. More concretely, this chapter aims at analysing the internal process of manual text simplification from two different points of view: the temporal and cognitive cost of manual simplification, and the difficulties that concrete rules pose to manual simplification, which will help determining future priorities for implementation. Next, Section 7.2 will present the "*Text Simplification Task Experiment"*.

7.2. General Description of the Text Simplification Task Experiment

This section contains the general description of the *"Text Simplification Task Experiment"*. Section 7.2.1 will present the general setting, while Section 7.2.2 will provide the numerical data relative to the texts used.

7.2.1. General setting of the experiment

The experiment described in this chapter consisted of asking six linguists—English advanced and native speakers with a Computational linguistics background—to carefully read and familiarise themselves with the CLCM guidelines (described in Section 4.3) and to manually simplify four texts of a total of two thousand words according to the simplification rules in the guidelines (Section 4.3.2). Even though the CLCM guidelines were provided, the linguists were left to their own judgement to decide which rules to apply and which not. In order to direct the participants and simplify the task of remembering over eighty rules, an assisting document was provided. The document (provided in Appendix B) contained fields to be filled in during the experiment and a list of the thirty most important rules, to be consulted during simplification.

The rules in the list were classified into the three natural language generation groups, as described in Section 4.3.2, namely:

- Rules for the discourse structure organisation at text level—those which determine and structure the content and thus act on the macro-discourse structure of the text (information presentation, order, and grouping).

- Rules for the discourse structure organisation at paragraph level—those acting at the level of micro-discourse structure—not at the level of the whole text, but at the level of separate sections or paragraphs of it.

- Concrete linguistic realisation rules—those acting at the sentence level and affecting concrete realisation choices at the word and sentence level.

The experiment was performed in two stages, distributed evenly over the course of two days to avoid the impact of the factor of fatigue. After completion of the entire simplification experiment, the participants were asked to complete a questionnaire (described in more detail in Section 7.4) and to provide feedback about the CLCM guidelines and the completed simplification work.

The texts used for the experiment were taken from the Crisis Management Corpus (described in Chapter 3). Specifically, they are from the Centres for Disease Control and Prevention[43] documents and represent instructions for the general population in different emergency situations. Table 7.1 shows the texts' distribution over the course of the two days, the texts' topics, and their respective lengths in words and characters.

Day	Text			
Day 1	Text 1	Returning Home after a Disaster. Clean Your Home and Stop Mold.	166 words, 900 chars	999 words
	Text 2	After a Flood. Precautions When Returning to Your Home.	833 words, 5018 chars	
Day 2	Text 3	Fact Sheet. Facts About Personal Cleaning and Disposal of Contaminated Clothing.	271 words, 1562 chars	999 words
	Text 4	Fact Sheet. Key Facts About Protecting Yourself After a Volcanic Eruption.	728 words, 4486 chars	

Table 7.1: Texts used in the CLCM guidelines internal evaluation.

As can be seen in Table 7.1, the first column divides the days into Day 1 and Day 2, with texts 1 and 2 (second column) shown in Day 1 and texts 3 and 4 shown in Day 2. The titles of the texts can be seen in the third column. As mentioned before, the total length of texts per day (fourth column) was held constant and consisted of around one thousand words. In order to keep the analysis of the experiment under control, the difficulty of the original texts was kept similar, in order to not have to take into account another text variable. The comparable difficulty of the four texts was again ensured by their provenance from the same source document as for the texts in Chapter 6. The

43 CDC, http://www.cdc.gov/, last accessed on March 9th, 2012.

results of the numerical analysis of the original texts and their manually simplified versions are provided in Section 7.2.2.

7.2.2. Quantitative description of the texts used

As a result of the experiment, six simplified versions of each original text in total (one per each linguist) were obtained, leading to four original texts and twenty-four simplifications. This section presents a comparison between the original and simplified versions of each text in terms of (1) text length (presented in Table 7.2) and (2) text complexity (Table 3).

Text/ Linguist	Original	Linguist 1	Linguist 2	Linguist 3	Linguist 4	Linguist 5	Linguist 6	Mean	Stand. dev.
Text 1	165	**176**	149	**226**	**231**	**212**	158	**192**	35.60
Text 2	771	708	214	702	651	**843**	751	645	222.66
Text 3	267	254	208	242	252	**288**	249	249	25.63
Text 4	575	392	225	498	459	**705**	**683**	494	181.24

Table 7.2: Length in words of the complex texts and their simplified versions.

In Table 7.2, the rows contain the length in words of each text, while the columns contain the respective values for each linguist. The last two columns contain the average values of the simplified text for all linguists with their standard deviations. The lengths of any simplified versions which are longer than the original texts are shown in bold. As can be seen, most of the simplified versions have a smaller number of words than the original texts, except for Text 1, which has longer simplified versions for four out of six simplified texts. It can also be seen that all of the simplifications of Linguist 5 are longer than the original texts. In addition to measuring the texts' length, a text complexity (TC) analysis similar to the one run in Chapter 3 was conducted. The analysis employed the same Python scripts as those described in Section 3.2.2. The TC features which were analysed were a subset of those which were analysed in Chapter 3 and can be regarded

as divided into *Main high TC issues* and *Secondary high TC issues*, namely:

1. *Main high TC issues:* Average sentence length (ASL), Average word length (AWL), Lexical diversity (LD), Average number of word senses (ANWS), and Proportion of function words (PFW).

2. *Secondary high TC issues:* Proportion of coordination markers (PCM), Proportion of subordination markers (PSM), Proportion of relative clause markers (PRCM), Proportion of ambiguous quantifiers (PAQ), and Proportion of personal and possessive pronouns (PPPP).

In comparison with the original set of high TC issues analysed in Chapter 3, only two of them were not evaluated in the texts in this experiment: the Proportion of punctuation signs and the Proportion of discourse markers. Specifically, the Proportion of punctuation signs (PPS) was not taken in consideration as a marker of high TC. This was done because, as can be seen in the simplified texts in Appendix E, they are divided into smaller elements, each of them delimited by a punctuation mark (“,”, “.”, “:”), and thus contain much more punctuation than the original versions of the text. The texts were also not analysed for incidence of discourse markers, since at the time of the experiment, the corresponding rule was not part of CLCM. The numerical incidence of the high TC markers which were measured is shown in Table 7.3.

High TC issues	**Original text**	**Simplif. 1**	**Simplif. 2**	**Simplif. 3**	**Simplif. 4**	**Simplif. 5**	**Simplif. 6**	**Average 1-6**
Main high TC issues								
ASL	*15.922*	10.169 √	9.636 √	11.206 √	10.352 √	10.893 √	12.165 √	**10.799** √
AWL	*5.327*	5.126 √	5.342	5.485	5.538	5.378	5.425	**5.389**
LD	*0.976*	0.980	0.932 √	0.964 √	0.963 √	0.950 √	0.968 √	**0.961** √
ANWS	*8.478*	8.848	9.660	8.958	8.833	8.478	8.446 √	**8.811**
PFW	*0.409*	0.347	0.361	0.340	0.312	0.359	0.380	**0.349**
Secondary high TC issues								

PCM	*0.063*	0.037 √	0.036 √	0.033 √	0.026 √	0.035 √	0.049 √	**0.036** √
PSM	*0.040*	0.039 √	0.039 √	0.041	0.040	0.037 √	0.035 √	**0.039** √
PRCM	*0.014*	0.011 √	0.009 √	0.010 √	0.005 √	0.012 √	0.008 √	**0.009** √
PAQ	*0.010*	0.004 √	0.003 √	0.005 √	0.003 √	0.009 √	0.007 √	**0.005** √
PPPP	*0.048*	0.036 √	0.052	0.040 √	0.047 √	0.037 √	0.049	**0.042** √

Table 7.3: Comparative TC analysis of the original and simplified texts.

The rows of Table 7.3 correspond to the high TC issues, while the columns show the values corresponding to the original texts and their simplifications. The columns from three to nine contain the values obtained for the four simplified texts produced by each linguist. The final column contains the average values of all twenty-four simplifications altogether. The values in the columns from three to nine (representing the simplifications) have to be compared with the values in column two (*Original text*). The positive impact of the CLCM simplification for each TC issue is marked in each simplified version cell by a "√", while the negative impact and no impact are not marked at all. As can be seen, most of the signs of the Secondary high TC issues are "√" and several of the Main high TC issues. As explained in Chapter 3, if the text simplification method had a positive impact, the values for all of the high TC issues, except for PFW, should be lower for the simplifications than for the original text. Since a high proportion of function words corresponds to a low number of content words, and thus less lexical richness, then if the simplification has a positive impact on complexity, the relationship between the values of the simplified texts and of the original should be inverse (i.e., the simplified texts should have a higher value than the original one). Statistical significance was calculated for the comparison between the original text values and the average of all simplifications and shows that all of the differences are significant at the 95% confidence level. Except for the Proportion of subordination markers and the Proportion of ambiguous quantifiers, where, although the results are positive, the small size of the sample prevents from having high confidence. The negative impacts on *Average word length* and *Average number of word senses* can be explained by the fact that the linguists did not have access to any resources which would allow them to find an appropriate synonym. In summary, the CLCM simplification has a positive impact,

and the simplified texts exhibit lower levels of text complexity than the original ones.

7.3. Evaluating the Manual Simplification Cost

This section will describe the evaluation of the cost involved in manually simplifying texts. The research hypothesis tested in this section is that manually simplifying texts according to the CLCM guidelines requires only low manual labour cost. The hypothesis will be tested by measuring the time taken for manual simplification (Section 7.3.1) and by comparing the different simplified versions of the same original text (Section 7.3.2).

7.3.1. Measuring the time taken to simplify the texts

As described in Section 7.2.1, the Text Simplification Task Experiment was divided into two days, in order to avoid the effect of fatigue. At the beginning of the first day, the participants had to read and get acquainted with the CLCM guidelines, and then provide feedback about how much time it took them to read the CLCM guidelines for the first time. The average time for this first, initial reading of the guidelines was between thirty and forty-five minutes. In addition, each day the participants received a sheet of paper containing as a reminder the thirty most-used rules, together with fields in which they had to fill in the time taken to simplify each of the two texts. Figure 7.1 shows the filled-in fields for the first day for Linguist 1.

2. Measure the time needed for simplification. If you need to interrupt – stop measuring the time.

Text:	Time:
Text1-166-words-protect-from-mold.txt	41 minutes
Text2-833-words-after-flood.txt	1 hour 12 minutes

Figure 7.1: Filled-in times for Day 1 of Simplification 1.

The screenshot in Figure 7.1 shows the part of the assisting document (provided in Appendix E) filled in by Linguist 1. The results of measuring the time employed for simplifying each text by each linguist are shown in Table 7.4.

Text/Linguist	Linguist 1	Linguist 2	Linguist 3	Linguist 4	Linguist 5	Linguist 6
Text 1	41 minutes	12 minutes	30 minutes	30 minutes	48 minutes	27 minutes
Text 2	72 minutes	14 minutes	105 minutes	60 minutes	96 minutes	69 minutes
Text 3	22 minutes	9 minutes	20 minutes	16 minutes	46 minutes	23 minutes
Text 4	22 minutes	17 minutes	30 minutes	37 minutes	93 minutes	39 minutes

Table 7.4: Comparison of the time taken to simplify the texts by the six linguists.

Table 7.4 lists the texts on the rows, and the linguists in its columns. As the original texts differed in length (as was seen in Tables 7.1 and 7.2), the time was normalised per length of the original texts in characters, thus obtaining the simplifying speed. Table 7.5 shows the results obtained in this way. The values are given in characters/minute and are rounded to the first digit after the decimal sign.

Linguists/Texts	Text 1	Text 2	Text 3	Text 4
Linguist 1	21.9	69.7	71.0	203.9
Linguist 2	75.0	358.4	173.6	263.9
Linguist 3	30.0	47.8	78.1	149.5
Linguist 4	30.0	83.6	97.6	121.6
Linguist 5	18.7	52.3	33.9	48.2
Linguist 6	33.3	72.7	67.9	115.0
Mean six	**34.8**	**114.1**	**86.5**	**150.3**
Stand. Dev. six	18.7	110.0	43.0	68.7

Mean five	**26.8**	**65.2**	**69.7**	**127.6**
Stand. Dev. five	5.5	13.3	20.7	50.6

Table 7.5: Simplifying speed per linguist.

Unlike Table 7.4, Table 7.5 lists the speeds per text in the columns and per linguist on the rows. A higher value (higher speed) indicates that the text was easier to simplify, while a lower value (lower speed) indicates that it was more difficult to simplify. It can be seen that the speed varied from 18.7 characters/minute (the lowest, by Linguist 5) for Text 1 to 358.4 characters/minute (the highest, by Linguist 2) for Text 2. If characters/minute are transformed into words/minute, taking into account that the average word length in the Crisis Management Corpus is 5.462 ± 0.005 (as stated in Table 3.5 in Section 3.3.1), this would be between 3.42 and 65.61 words per minute.

It can also be noted that the simplifying speeds of Linguist 2 (values underlined) were very high compared to the other participants. For this reason, this participant is considered to be an outlier.

Mean value and standard deviations per text for all of the six linguists and mean + standard deviation for all of the linguists without the outlier are listed in the last four rows. If the outlier is removed, the means for the remaining five linguists show that the speeds are increasing from Text 1 to Text 4. This can be explained by the fact that the linguists learn to simplify better and that there is a learning effect. Due to the learning effect, the average speed was calculated, taking into account only the time results for the fourth text and removing the outlier. This result is 127.6 characters per minute (i.e. 23.36 words per minute). Although this low manual simplification speed indicates that manually simplifying texts according to the CLCM guidelines requires a high manual labour cost and thus does not support the research hypothesis, the learning effect shows that with more training the linguists may speed up their simplification work.

7.3.2. Comparing the concrete simplifications

An additional analysis of aligned simplified versions of Text 1 (Table 7.6 and Table 7.7) showed that linguists produced simplified versions of the same original text which were different both in terms of text unit ordering and in terms of rendering the same text unit. Table 7.6 presents a comparison of the complex Text 1 with two different simplifications (2 and 6).

Order	Original Text 1	Order	Simplification 2	Order	Simplification 6
1, title	Clean Your Home and Stop Mould	**1, title (Original 1)**	How to clean your home and stop mould	**1, title (Original 1)**	Clean Your Home and Stop Mould
				2, Subsection title	Water Leaks
2	Take out items that have soaked up water and that cannot be cleaned and dried.	**2 (Original 2)**	If there are water-soaked items AND If there are items that cannot be clean and dried: Take out items.	**3 (Original 3)**	Fix water leaks
3	Fix water leaks.	**3 (Original 3)**	Fix water leaks.	**4 (Original 2)**	Take out soaked items you cannot clean and dry.
4	Use fans and dehumidifiers and open doors and windows to remove moisture.	**4 (Original 4)**	Use fans and dehumidifiers.	**5 (Original 4)**	Use fans and dehumidifiers.
5	To remove mould, mix 1 cup of bleach in 1 gallon of water, wash the item with the bleach mixture, scrub rough surfaces with a stiff brush, rinse the item with clean water, then dry it or leave it to dry.	**5 (Original 4)**	Open doors and windows.	**6 (Original 4)**	Open doors and windows to remove moisture.
6	Check and clean heating, ventilating, and air-conditioning systems before use.	**6 (Original 5)**	How to remove mould:	**7, Subsection title**	Warning
7	To clean hard surfaces that do not soak up water and that may have been in contact with	**7 (Original 5)**	Mix 1 cup of bleach in 1 gallon of water.	**8 (Original 12)**	Never mix bleach and ammonia!

	floodwater, first wash with soap and clean water.				
8	Next disinfect with a mixture of 1 cup of bleach in 5 gallons of water.	**8 (Original 5)**	Wash the item with the bleach mixture.	**9 (Original 13)**	Vapour from mixture kills.
9	Then allow to air dry.	**9 (Original 5)**	Scrub rough surfaces with a stiff brush.	**10, Subsection title**	To Remove Mould
10	Wear rubber boots, rubber gloves, and goggles when cleaning with bleach.	**10 (Original 5)**	Rinse the item with clean water.	**11 (Original 5)**	Mix 1 cup of bleach in 1 gallon of water.
11	Open windows and doors to get fresh air.	**11 (Original 5)**	Dry the item or leave the item to dry.	**12 (Original 5)**	Wash the item with the bleach mixture.
12	Never mix bleach and ammonia.	**12 (Original 6)**	Check and clean heating systems.	**13 (Original 5)**	Scrub rough surfaces with a stiff brush.
13	The fumes from the mixture could kill you.	**13 (Original 6)**	Check and clean ventilating systems.	**14 (Original 5)**	Rinse the item with clean water.
		14 (Original 6)	Check and clean air-conditioning systems.	**15 (Original 5)**	Dry it OR Leave it to dry.
		15 (Original 7)	How to clean hard surfaces:	**16, Subsection title**	Necessary Checks
		16 (Original 7)	Wash with a mixture of soap and clean water.	**17 (Original 6)**	Before use, check and clean: heating systems, ventilating systems, air-conditioning systems.
		17 (Original 8)	Disinfect with a mixture of 1 cup of bleach in 5 gallons of water.	**18, Subsection title**	To Clean Hard Surfaces
		18 (Original 9)	Allow to air dry.	**19 (Original 7)**	If hard surfaces do not soak up water AND May have been in contact with floodwater, Wash with soap and clean water.
		19 (Original 10)	When cleaning with ammonia: Wear rubber boots. Wear rubber gloves Wear goggles.	**20 (Original 8)**	Next disinfect with a mixture of 1 cup of bleach in 5 gallons of water.
		20 (Original 11)	Open windows and open doors.	**21 (Original 10)**	When cleaning with bleach, wear: rubber boots, rubber gloves, goggles.

		21 (Original 12)	Never mix bleach and ammonia.	**22 (Original 9)**	Allow to air dry.
		Original 13	MISSING	**23 (Original 11)**	Open windows and doors to get fresh air.

Table 7.6: Comparison of a complex text and two simplifications.

Table 7.6 shows the order of sentences in Original Text 1 and its Simplifications 2 and 6 (i.e. the simplifications produced by Linguists 2 and 6). Columns 1, 3, and 5 display the serial number of the sentence in the respective text, with column 1 showing the number of the sentences in Original Text 1, column 2 showing the number of the sentences in Simplification 2, and column 5 showing the number of the sentences in Simplification 6. Next to each number of each sentence in each simplification is given the number of the corresponding sentence in the original text. It can be seen that Simplification 2, although splitting the original sentences into smaller units, mostly follows the same order as Original Text 1. In fact, the numbers of its sentences correspond to the following numbers of sentences in the original text: 1, 2, 3, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9, 10, 11, 12. Simplification 6, instead, has implemented some re-ordering—specifically, it has inserted many more titles of sub-sections, and the corresponding numbers of its sentences in the original text are in the following order: 1, 3, 2, 4, 4, 12, 13, 5, 5, 5, 5, 5, 6, 7, 8, 10, 9, 11.

Table 7.7 presents an aligned comparison of the same segments in three different simplifications for part of Text 1. This is done in order to assess whether the same CLCM rules were applied to the same text complexity issue. Column 1 lists the original segment in the complex text, while the next columns show its rendering in the three simplifications. Next to each change applied by each linguist the relevant concrete CLCM rules are shown.

	original	**simplification 1**	**simplification 2**	**simplification 3**
title	Clean Your Home and Stop Mold	How to clean your home and stop mold In_T_S_01	How to clean your home and stop mold In_T_S_01	How to clean your home and stop mold In_T_S_01

e				
1	Take out items that have soaked up water and that cannot be cleaned and dried.	Remove: items that have soaked up water AND items that cannot be cleaned and dried. In_S_09	If there are water-soaked items AND If there are items that cannot be clean and dried: Take out items. In_Cd_G_02 In_Cd_P_01 In_Cd_S_01	Remove from home: Wet items, Items that cannot be cleaned and dried. In_P_03 In_Li_P_01 In_Li_P_02 In_S_11
2	Fix water leaks.	Fix water leaks.	Fix water leaks.	Fix water leaks.
3	Use fans and dehumidifiers and open doors and windows to remove moisture.	To remove moisture: Use fans. Use dehumidifiers. AND Open doors AND windows. In_P_03 In_I_G_02 In_S_09 In_Cd_P_01 In_I_F_01	Use fans and dehumidifiers. Open doors and windows. In_I_G_01 In_I_G_02	Use fans AND dehumidifiers. In_S_09
4	To remove mold, mix 1 cup of bleach in 1 gallon of water, wash the item with the bleach mixture, scrub rough surfaces with a stiff brush, rinse the item with clean water, then dry it or leave it to dry.	To remove mold: Mix: 1 cup of bleach AND 1 gallon of water. Wash the item with the bleach mixture. Use a stiff brush to clean rough surfaces on the item. Rinse the item with clean water. Dry the item. OR Leave the item to dry. In_Cd_P_01 In_S_09 In_I_G_01 In_I_G_02 In_Cd_G_03 In_I_F_01	How to remove mold: Mix 1 cup of bleach in 1 gallon of water. Wash the item with the bleach mixture. Scrub rough surfaces with a stiff brush. Rinse the item with clean water. Dry the item or leave the item to dry. In_T_S_01 In_I_G_01 In_I_G_02 In_I_F_01	If items have mold: Remove mold from items. How to remove mold from items: Mix 1 cup of bleach AND 1 gallon of water. Wash the item with the bleach mixture. If the item has rough surfaces: Scrub the rough surface with a stiff brush. Rinse the item with clean water. Dry the item OR Leave them item to dry. In_Cd_P_01 In_Cd_S_01 In_I_F_01 In_I_G_04

Table 7.7: Elements' alignment of Text 1 with three simplifications.

As can be seen from Table 7.7, except for the title, which was rendered in the same way, and the short sentence “Fix water leaks.”, which remained unchanged, all of the other segments were rewritten in completely different ways. Although some of the rules are repeated (highlighting indicates repetition), Table 7.7 also shows that different sets of rules were applied by different linguists to simplify the same text unit. This suggests that there are situations in which the choice of which simplification rule to apply is not unique, requiring the simplifier to make a decision, and it is considered that this may increase the cognitive load on the linguists.

It is hypothesized that if the linguists’ agreement were calculated, it would be very low, because of the large differences in the alternative simplifications. It is suggested that implementing part of these simplification operations and presenting the linguist with an already partially rewritten text may reduce both the cognitive and the temporal effort of a linguist carrying out a simplification.

7.4. Evaluating the Difficulty of Applying Concrete Simplification Rules

As mentioned in Section 7.2.1, at the end of the second day, the participants were asked to fill in a questionnaire eliciting details regarding the work which they had done in the previous two days. The questionnaire collected data in three parts: Part 1 asked for feedback in the form of free text, Part 2 provided a list of rules to be evaluated in terms of how difficult they are to apply to produce manual simplifications, and Part 3 suggested a list of assistive automatic implementations to be rated. This section will treat the setting and the results collected on the basis of the first two parts of the questionnaire (Part 1 in Section 7.4.1 and Part 2 in Section 7.4.2), while the next section will discuss the data collected on the basis of Part 3 of the questionnaire.

7.4.1. Free text feedback results from the questionnaire

As can be see in Appendix E, the feedback in the first part of the questionnaire was elicited by the following question: “Could you think of what was most difficult for you while simplifying?” The responses of the linguists consisted of enumerating concrete simplification operations with which they had problems. The following responses were given:

- *Being mindful of word order and word difficulty.*
- *Avoiding negatives was difficult/rephrasing negative phrases.*
- *Avoiding relative clauses was difficult. The guidelines said to avoid relative clauses, but didn’t suggest an alternative.*
- *Knowing what to include and what could be left out was difficult.*
- *Following the rules for avoiding present participles was difficult.*
- *Remembering to remove pronouns*
- *Re-organizing and regrouping the content of the original/grouping together different items*
- *Deciding whether the original text is an explanation, a conditional, or actually an instruction.*
- *Ordering instructions in chronological order/making decisions about chronological order*
- *Complex conditions, e.g. if A OR if B AND C*
- *Making decisions about how to rewrite conditions*
- *Remembering the Message instructions*
- *Dealing with lists inside comments/explanations*
- *Writing the explanations without using (demonstrative) pronouns*
- *Dealing with lists of references/websites*
- *Avoiding ambiguous words*

- *Replacing technical words*
- *Avoiding negation*
- *Repeating complements and adjectives without changing the meaning of the sentence*

Note that certain families of responses were given more than once, such as those pertaining to negation. To quantify this, responses were grouped into categories, counted, and averaged. Figure 7.2 shows their relative ranking.

Avoiding negatives/Rephrasing negative phrases.	0.5
Remembering to remove pronouns./Avoiding pronouns.	0.5
Being mindful of word difficulty/Replacing technical words.	0.3333333333
Re-organizing and regrouping the content of the original.	0.3333333333
Being mindful of word order.	0.1666666667
Avoiding relative clauses.	0.1666666667
Knowing what to include and what could be left out.	0.1666666667
Following the rules for avoiding present participles.	0.1666666667
Deciding if original text was an explanation, a conditional, or actually an instruction.	0.1666666667
How to order instructions in chronological order.	0.1666666667
Grouping together different items.	0.1666666667
Making decisions about how to rewrite conditions/ Complex conditions eg. If A OR If B AND C.	0.1666666667
Remembering the Message instructions.	0.1666666667
Writing comments/explanations etc. that contain lists.	0.1666666667
Understanding which rules are applicable to comments.	0.1666666667
Knowing what to do with lists of references/websites etc. for further information.	0.1666666667
Avoiding ambiguous words.	0.1666666667
Repeating complements and adjectives without changing the meaning of the sentence.	0.1666666667

Figure 7.2: Relative ranking of responses in Part 1 of the questionnaire.

Figure 7.2 lists the concrete simplification problems on the left and the average scores on the right. The first four simplification operations were the most frequently cited, namely:

1. Rephrasing negative phrases and replacing pronouns with their antecedents (both ranked at the highest position), and

2. Replacing technical words with common synonyms and re-organizing the content of the original text (both ranked at the second position).

All of the other problems have the same average weight. Next, Section 7.4.2 will provide a more concrete analysis of the CLCM simplification rules.

7.4.2. Concrete rule evaluation results from the questionnaire

Part 2 of the questionnaire provided a list of rules to be evaluated in terms of how difficult they are to apply to produce manual simplifications. The list was taken from the assisting document (described in Section 7.2.1) and thus represented the thirty most important rules to be consulted during simplification. As can be seen in Appendix E, the linguists were asked to mark each of the listed rules as “difficult” or “easy”, and to indicate whether automatically implementing this rule would “simplify” or speed up their simplification work. Table 7.8 shows the responses of the six linguists. The rules are formulated as displayed in the assisting document.

Rule/Linguist	Linguist 1	Linguist 2	Linguist 3	Linguist 4	Linguist 5	Linguist 6
identify the separate situations 2	easy	easy	difficult	difficult, simplify	simplify	easy
group information regarding the specific situations in separate blocks 2	easy	easy	difficult	difficult, simplify	simplify	easy
jump two new lines after every specific situation block 3	easy, simplify	easy	easy	easy simplify	easy, simplify	easy
provide a unequivocal title for each specific situation block 3	easy	difficult, simplify	easy	difficult, simplify	simplify	easy
use the allowed formulations for the titles 3	easy	moderate, simplify	easy	difficult, simplify	simplify	easy
jump two new lines after	easy,	easy	easy	easy	easy,	easy

each title 3	simplify			simplify	simplify	
order instructions in logical and chronological order 2	difficult	easy	difficult	difficult, simplify	difficult, simplify	easy
place conditions before instructions 3	easy	easy	easy	difficult, simplify	difficult, simplify	difficult, simplify
use standard word 4 order	difficult, simplify	easy	easy	difficult, simplify	simplify	difficult, simplify
use the suggested formulations for conditions 2	easy	easy	easy	easy, simplify	simplify	easy
if you coordinate two conditions - write one on one line, then "AND"or "OR" and the other one on the second line 4	easy, simplify	easy, simplify	easy	difficult, simplify	easy, simplify	easy
put the more specific conditions before the more general ones 2	difficult	easy	easy	difficult, simplify	difficult, simplify	easy
if there are two actions to be done simultaneously, write: "Do these two actions simultaneously:" 3		easy, simplify	easy	easy, simplify	simplify	easy
order explanations, exceptions and other notes after instructions 3		easy	difficult	easy, simplify	simplify	easy
put a colon after a condition 3	easy	easy	easy	easy, simplify	easy, simplify	difficult, simplify
write only one action per line 4	easy, simplify	easy	easy	difficult, simplify	difficult, simplify	difficult, simplify
replace technical terms with common synonyms 3	difficult	easy, simplify	difficult	difficult, simplify	difficult, simplify	easy
replace idiomatic expressions with literal ones 3		difficult, simplify	difficult	difficult, simplify	difficult, simplify	easy
replace enumerations with vertical lists 4	easy, simplify	easy, simplify	easy	easy, simplify	simplify	easy
write the cardinal numbers in figures 4	easy, simplify	easy	easy	easy, simplify	simplify	difficult, simplify
expand the abbreviations at their first occurrence 2		easy	easy	easy, simplify	simplify	easy
avoid any pronouns (personal, possessive, demonstrative) 4	difficult, simplify	difficult, simplify	easy	difficult, simplify	difficult, simplify	easy
avoid ambiguous words 3	easy	difficult, simplify	difficult	difficult, simplify	difficult, simplify	easy
keep the preposition and the verb together in phrasal verbs 3	difficult	easy	easy	easy, simplify	simplify	difficult, simplify
replace passive with active voice 3	difficult	easy	difficult	difficult, simplify	difficult, simplify	difficult, simplify

try to avoid negative forms 4	difficult, simplify	very difficult, simplify	very difficult	difficult, simplify	difficult, simplify	easy
if a preposition/adjective refers to more than 1 noun, repeat the preposition/adjective next to each noun 4	difficult, simplify	easy	easy	difficult, simplify	simplify	difficult, simplify
if more than 1 complement determine the same noun, repeat the noun 4	difficult, simplify	easy	easy	difficult, simplify	simplify	difficult, simplify
put a comma after each element of a list, except of the last one (put a dot at the end of the last element of a list) 3	easy	difficult, simplify	easy	easy, simplify	simplify	easy

Table 7.8: Part 2 responses.

Table 7.8 lists the CLCM rules in the first column and the responses of each linguist in the following columns. Note that Linguist 3 did not provide any "simplify" suggestions.

To allow comparison of the different combinations of assessments of ease or difficulty and the utility of automatic assistance, the responses have been assigned weights in correspondence with the different linguists' ranks of the different marks. Since the purpose of this evaluation was to estimate the difficulty of the concrete rules, the weights assigned were focussed on the "easy"/ "difficult" responses, while the suggestion "simplify" has been taken into account as an additional weight to be added to the main rule weight. The scores were thus assigned in the following way:

- "no answer" or "easy" = 0
- "simplify" = 1
- "moderate" = 1.5
- "difficult" = 2

- “difficult” + “simplify” = 3
- “very difficult" = 4
- “very difficult” + “simplify” = 5

The weights of the responses were summed and average scores were obtained. Figure 7.3 shows the rule difficulty scores obtained from Part 2 of the questionnaire.

try to avoid negative forms	3
replace passive with active voice	2.17
avoid any pronouns (personal, possessive, demonstrative)	2
avoid ambiguous words	1.83
replace idiomatic expressions with literal ones	1.83
replace technical terms with common synonyms	1.83
order instructions in logical and chronological order	1.67
if more than 1 complement determine the same noun, repeat the noun	1.67
write only one action per line	1.67
if a preposition/adjective refers to more than 1 noun, repeat the preposition/adje	1.67
use standard word order	1.67
place conditions before instructions	1.5
put the more specific conditions before the more general ones	1.33
provide a unequivocal title for each specific situation block	1.17
keep the preposition and the verb together in phrasal verbs	1.17
identify the separate situations	1
group information regarding the specific situations in separate blocks	1
if you coordinate two conditions - write one on one line, then “AND”or “OR” a	1
write the cardinal numbers in figures	1
use the allowed formulations for the titles	0.92
put a comma after each element of a list, except of the last one (put a dot at the e	0.83
put a colon after a condition	0.83
order explanations, exceptions and other notes after instructions	0.67
replace enumerations with vertical lists	0.67
jump two new lines after every specific situation block	0.5
if there are two actions to be done simultaneously, write: “Do these two actions s	0.5
jump two new lines after each title	0.5
use the suggested formulations for conditions	0.33
expand the abbreviations at their first occurrence	0.33

Figure 7.3: Rule difficulty score results from Part 2 of the questionnaire.

Similarly to Figure 7.2, Figure 7.3 lists on the left the concrete CLCM rules and on the right their average marks. It can be seen that the most difficult to apply rules are:

1. “try to avoid negative forms”, ranked first
2. “replace passive with active voice”, ranked second
3. “avoid any pronouns” , ranked third, and
4. “avoid ambiguous words”, “replace idiomatic expressions with literal ones” and “replace technical terms with common synonyms”, all ranked fourth.

The rules ranked as the least difficult were:

1. “expand the abbreviations at their first occurrence” and “use the suggested formulations for conditions”, both ranked last.
2. “jump two new lines after each title”, “if there are two actions to be done simultaneously, write ...” and “jump two new lines after each specific block”, all ranked penultimately.

Note that the most difficult rules are those which involve comprehending and making a change to a text complexity issue which bears a high cognitive load, while the rules which are considered to be the easiest ones mostly involve formatting or a change applied only once per document. The results obtained on the basis of Part 2 of the questionnaire can be used for future work investigating whether these rules were formulated in an easy-to-understand way, and for determining priorities for future implementation of text simplification assistive systems. Next, Section 7.5. will analyse the suggestions given by the linguists concerning what would be preferable to be implemented automatically in order to assist them in their simplification work.

7.5. Investigating Implementation Priorities

This section will present the investigation of the priorities for future implementation of text simplification assistive systems and its motivations. Section 7.5.1 will present the motivations,

while Sections 7.5.2 and 7.5.3 present the two methods of investigation.

7.5.1. Motivations for the investigation

Due to the amount of time taken by linguists to manually simplify emergency instructions, which, as stated in Section 7.3, was on average 23.36 words per minute for the five linguists for the fourth text, and the extensive feedback received by the linguists regarding which rules pose the most problems, it seems clear that a text simplification assistive tool would be useful. This finding suggests the utility of investigating the implementation priorities for a future semi-automatic simplification assistive tool. The choice to proceed with the implementation of a new semi-automatic tool was based on the conclusions in Section 2.4 that:

Both the existing fully automatic text simplification systems and semi-automatic tools address a very limited set of high text complexity issues, which is probably due to difficulty of implementation and evaluation.

The existing fully- and semi-automatic text simplification systems are focussed on a set of simplification operations which are not tailored for the crisis management domain.

For these reasons, and due to the fact that the incorrect comprehension of written documents in the crisis management domain can lead to substantial loss of lives, money, and property, it is considered that it would be appropriate to implement a semi-automatic text simplification tool. The text simplification of crisis management documents with this tool will thus consist of computer-aided text simplification, in which some operations will be performed automatically by the tool, and then revised, and the simplification completed manually, by end-users. Such an implementation would

both surpass the limitations of the extant fully automatic simplification systems and assist end-users in their simplification work.

7.5.2. Investigation of suggestions from Part 2 of the questionnaire

The investigation of implementation priorities was based on the 'simplify' responses in Part 2 of the questionnaire and the feedback which was collected from Part 3. This time, the results from Part 2 were considered with a focus on the number of "simplify" marks given to the CLCM rules by the linguists.

The rules with the highest number of "simplify" marks (four) were collected in a list and ranked according to the rule difficulty scores obtained in Section 7.4. The ranking obtained in this way is provided in Table 7.9, ordered from the hardest rule to the easiest rule.

Rule	N. of 'simplify'	Difficulty score
• Try to avoid negative forms.	4	3
• Avoid any pronouns (personal, possessive, demonstrative).	4	2
• Use standard word order. • Write only one action per line. • If a preposition/adjective refers to more than 1 noun, repeat the preposition/adjective next to each noun. • If more than 1 complement determines the same noun, repeat the noun.	4	1.67
• If you coordinate two conditions - write one on one line, then "AND"or "OR" and the other one on the second line. • Write the cardinal numbers in figures.	4	1
• Replace enumerations with vertical lists.	4	0.67

Table 7.9: CLCM rules ranked according to the most "simplify" suggestions.

As can be seen, Table 7.9 lists the rules in the left column, the number of "simplify" responses in

the middle column, and the average rule difficulty score (taken from Figure 7.3) in the right column. It can also be seen that there are four rules with an average score of 1.67 and two rules with a score of 1. It can be seen that the rules for which implementation was suggested vary from very cognitively difficult ones ("try to avoid negative terms" and "avoid any pronouns") to light formatting and replacement rules, such as "Write the cardinal numbers in figures." and "Replace enumerations with vertical lists.". Since Linguist 3 did not give any "simplify" suggestions, and due to the programming limitations already discussed in Section 2.1.3, it was necessary to study the preferences of the linguists regarding a set of proposed implementations.

7.5.3. Investigation of suggestions from Part 3 of the questionnaire

The second part of the investigation of future implementation directions consisted of the results collected from Part 3 of the questionnaire. As can be seen in the questionnaire in Appendix E, the linguists were provided with a list of suggested automatic implementations and asked to rate them in terms of how much their implementation would "simplify and speed up" their simplification work. The suggested implementations offered preliminary easy-to-implement operations which would result in highlighting different text elements. The text elements to be highlighted ranged from single words to whole paragraphs and were CLCM-specific. The reasons behind offering these operations are given in Table 7.10.

Proposed implementation	**Motivation**
The text is presented to you with highlighted separate thematic situations.	This would help by splitting text into separate thematic blocks, as according to rules **In_G_03, In_G_09**, **and In_G_10.**
The text is presented to you with highlighted acronyms and abbreviations.	To identify abbreviations in order to ease expanding them at their first encounter, according to the rule stated in the assisting document or to use the ones defined in the CLCM guidelines as stated in rule **In_L_06**.
The text is presented to you with highlighted pronouns.	In order to facilitate identifying pronouns in order to replace them with their antecedents (rules **In_S_02, In_S_03** and

	In_S_04.).
The text is presented to you with highlighted passive voice.	In order to facilitate identifying passive voice in order to replace it with active voice (rule **In_S_05**).
The text is presented to you with highlighted negative phrases.	In order to facilitate identification of negative phrases to replace them with positive expressions (rule **In_S_06**).
The text is presented to you with highlighted nouns to which a preposition refers.	In order to ease applying rule **In_S_12.**
The text is presented to you with highlighted nouns to which an adjective refers.	In order to ease applying rule **In_S_13.**
The text is presented to you with highlighted technical terms.	In order to facilitate the recognition of technical terms in order to replace them with more common words (rules **In_L_02,** and **In_L_04**).
The text is presented to you with highlighted and underlined verbs.	In order to ease applying the rule to keep one action per sentence (rule **In_I_G_02**).
The text is presented to you with highlighted beginning of conditions.	In order to assist with recognizing and rewriting conditions (rules **In_F_04, In_P_02, In_P_03, In_Cd_P_01, In_Cd_S_01,** etc.).
The text is presented to you with highlighted beginning of instructions.	In order to assist with identification and rewriting of instructions (rules **In_G_09**, **In_F_03, In_P_02, In_P_03, In_Cd_G_04, In_I_F_01, In_I_G_01, In_I_G_02, In_I_G_03** etc.).
The text is presented to you with highlighted beginning of explanations.	In order to ease recognizing and splitting the explanation, as a less important element (rule **In_Cm_G_01**).
The text is presented to you with highlighted whole conditional expressions.	In order to assist with recognizing and rewriting conditions (rules **In_F_04, In_P_02, In_P_03, In_Cd_P_01, In_Cd_S_01,** etc.).
The text is presented to you with highlighted whole instructions.	In order to assist with identification and rewriting of instructions (rules **In_G_09**, **In_F_03, In_P_02, In_P_03, In_Cd_G_04, In_I_F_01, In_I_G_01, In_I_G_02, In_I_G_03** etc.).
The text is presented to you with highlighted whole explanations.	In order to ease recognizing and splitting the explanation, as a less important element (rule **In_Cm_G_01**).
The text is presented to you with highlighted phrasal verbs in case the main verb and the preposition are split up.	In order to ease recognition of phrasal verbs in order to keep their main body and particle together, according to rule **In_L_05**.
The text is presented to you with ambiguous lexical terms highlighted.	In order to ease recognition of ambiguous lexical items and their replacement with less ambiguous synonyms (as stated in the Lexical rules section of the CLCM Guidelines).
The text is presented to you with ambiguous syntactic expressions highlighted.	In order to ease recognition of ambiguous syntactic expressions and their replacement with less ambiguous structures (as stated in the Forbidden syntactic structures section of the CLCM Guidelines).

Table 7.10: Proposed implementations and their motivations.

The participants were asked to give scores to these operations, according to the following ranking:

1 - Implementing this operation will not help at all.

2 - Implementing this operation will help to a certain extent.

3 - Implementing this operation will help very much.

The results from Part 3 of the questionnaire are given in Figure 7.4. The results were again summed and averages were ordered from the highest to the lowest one.

The text is presented to you with ambiguous lexical terms highlighted.	2.67
The text is presented to you with highlighted phrasal verbs in case the main verb and the pre	2.5
The text is presented to you with highlighted separate thematic situations.	2.5
The text is presented to you with highlighted negative phrases.	2.33
The text is presented to you with ambiguous syntactic expressions highlighted.	2.33
The text is presented to you with highlighted technical terms.	2.33
The text is presented to you with highlighted beginning of instructions.	2.17
The text is presented to you with highlighted beginning of conditions.	2.17
The text is presented to you with highlighted beginning of explanations.	2.17
The text is presented to you with highlighted acronyms and abbreviations.	2.17
The text is presented to you with highlighted pronouns.	2
The text is presented to you with highlighted passive voice.	2
The text is presented to you with highlighted nouns to which a preposition refers.	2
The text is presented to you with highlighted nouns to which an adjective refers.	1.83
The text is presented to you with highlighted whole conditional expressions.	1.83
The text is presented to you with highlighted whole instructions.	1.67
The text is presented to you with highlighted whole explanations.	1.5
The text is presented to you with highlighted and underlined verbs.	1.17

Figure 7.4: Ranking of proposed implementations in Part 3 of the questionnaire.

As can be seen in Figure 7.4, the left column contains the list of proposed implementation operations and the right column contains their average scores. Here again, the linguists prioritized the most difficult cognitive operations, such as ambiguous terms and negative phrases. In addition, some more cognitively light and easy to implement suggestions are ranked highly: to highlight phrasal verbs, in order to ease putting their main part and particle together. Next, Section 7.6 will summarize the findings presented in this section and in the previous Sections 7.3 and 7.4.

7.6. Summary of the Findings

The results presented in the previous sections of this Chapter can be summarized as follows:

- The simplified versions of the original texts, produced according to the CLCM rules, are not longer than their original versions.

- The text complexity (TC) levels of the simplified versions are lower in comparison with the original texts, according to several measures.

- The comparison of different simplified versions of the same original text showed that linguists produced different simplifications in terms of both the ordering of text units and the rendering of the same text unit.

- Although manual simplification requires a substantial amount of time, learning effect is visible. It was hypothesized that implementing part of these simplification operations and presenting the linguist with an already partially rewritten text may reduce both the cognitive and the temporal effort of the simplifying linguist.

- The results in Sections 7.4.1 and 7.4.2 led to the finding that the most difficult simplification rules to apply are those which affect processing of the most cognitively difficult TC issues, such as processing negation, passive voice, anaphora, ambiguity, and figurative language.

- The results in Sections 7.5.2 and 7.5.3 showed that the same cognitively difficult issues, which are at the same time the biggest challenges for NLP applications, are indicated as

preferable to be implemented automatically.

On the basis of the aforementioned findings (and the findings in Chapters 5 and 6), it can be concluded that CLCM-based manual text simplification produces good results in terms of text simplification, but is very time- and effort-intensive. This chapter also provides a list of functions to be considered first while implementing a computer-aided text simplification assistive tool. Next, Section 7.7 will provide the conclusions of this chapter.

7.7. Conclusions and Future Work

This chapter is the last one of the evaluation chapters, evaluating the CLCM simplification rules and guidelines which were presented in Chapter 4. After Chapters 5 and 6 showed a positive impact of the CLCM-based simplification on three tasks, the aim of this chapter was to evaluate how effective and how difficult it is to manually simplify texts according to this simplification approach and to draw conclusions about future implementation priorities. This investigation was based on a specially designed experiment (the *"Text Simplification Task Experiment"*), involving six computational linguists who were asked to manually simplify four texts of two thousand words altogether over a two-day period. The results of the experiment showed, as an additional achievement, that CLCM-based text simplification reduces the TC levels of emergency instructions. However, it also showed that the simplified versions of the same text which are thereby obtained are too diversified in terms of both the ordering of text units and concrete rendering of the text units and CLCM rules applied. This discovery, along with the finding that the time employed to simplify text is relatively long, leads to the conclusion that an at least partial automation of the simplification process would reduce the cognitive load of linguists in making the decision of which rules to apply. On the basis of this conclusion, a further investigation of the implementation priorities was

conducted. The analysis of the difficulty of application of concrete CLCM rules and the preferences of the linguists about which CLCM rules to implement automatically has led to a list of implementation priorities and to the unsurprising conclusion that the most difficult to apply rules for humans are those which are also the most challenging for NLP applications. The computational background of the participants has also, however, led to a few easier-to-implement rules. All of this insight will be used in future work involving the implementation of a computer-aided text simplification tool. Since “avoiding negatives” was listed as the first choice in Part 1 and Part 2 and also had one of the highest scores in Part 3, it can be considered to be the most urgent issue to be solved and possibly implemented. Negation implementation would include constructing patterns for recognizing negation to avoid in emergency instructions, based on the collected corpus and on building a grammar to assist in supplying the user with positive alternatives to negated phrases. Another candidate for implementation is, of course, “Highlighting the ambiguous lexical terms”, which emerged as the suggested implementation with the highest-ranking score (2.67). Future work would also include testing whether more appropriate training of human simplifiers would change which rules are considered difficult to apply. Next, Chapter 8 will provide the conclusions of this thesis.

Chapter 8 – Conclusions

This thesis provides original contributions to and addresses an important gap in the knowledge of language complexity and comprehensibility of Crisis Management written documents as well as of methods of rewriting these documents in simple and straightforward language. This chapter summarizes the results of the research presented in this thesis and its original contributions. Section 8.1 revisits the goals set out in Chapter 1 and discusses how they were achieved, Section 8.2 revisits the aims set out in Chapter 1 and summarizes the original contributions of this thesis achieved in its chapters, Section 8.3 reviews the contents of the thesis chapter by chapter, and finally, Section 8.4 provides directions for future work.

8.1. Thesis Goals Revisited

This Section revisits the goals set out in Chapter 1 and provides a description of how each goal was achieved.

Goal 1 was to identify and select a set of text complexity factors affecting the crisis management sub-language to be measured in the text complexity analysis and addressed in the writing guidelines for re-writing existing or producing new clear crisis management documents in English. The goal was achieved partially in Chapter 2 and completed in Chapter 3 on the basis of psycholinguistic literature related to reading comprehension and comprehension under stress, relevant NLP literature, and an overview of the existing approaches to measuring and reducing text complexity.

Goal 2 was to perform a critical review of the existing approaches to text complexity and text simplification and to investigate their applicability and/or their limitations with respect to documents in the Crisis Management domain. The goal was achieved in Chapter 2. The review analysed the existing approaches in terms of whether they are tailored for the crisis management domain and language and gave motivations regarding the final choice of both a text complexity measuring approach and a text simplification and clear documents writing approach.

Goal 3 was to collect data needed for the text complexity analysis of written documents in the Crisis Management domain. The goal was achieved in Chapter 3 with the collection of the Crisis Management Corpus. The corpus reflects the nature of communication in the crisis management domain and contains a representative set of documents addressing both of the main classes of readers (general population and specialists), covering a variety of sub-domains (general crisis management, aeronautics and medical) and of document types (instructions and alerts).

Goal 4 was to investigate the amount of high text complexity factors present in Crisis Management documents. The goal was achieved in Chapter 3. This involved conducting a text complexity corpus analysis of the corpus which was collected for achieving Goal 3. The corpus analysis required developing a set of Python scripts to address each particular high text complexity issue, as well as for calculating the statistical significance of the results. The results showed that all types of the crisis management documents studied exhibit a number of text complexity issues, which makes it necessary to apply a text simplification approach to them. The analysis also showed differences between the combinations of text complexity issues in different types of crisis management documents.

Goal 5 was to propose and develop an appropriate approach for writing and simplifying texts, based

on linguistic theory, which must be tailored to crisis management documents written in English. The goal was achieved in Chapter 4. The solution consisted of adapting one of the controlled language approaches presented in Chapter 2 from French to English by reflecting the particularities of English grammar. The proposed approach addresses all of the text complexity issues which were identified while achieving Goals 2 and 4.

Goal 6 was to perform an evaluation of the proposed approach for writing and simplifying texts in terms of whether it has a positive impact on the comprehensibility of emergency instructions. The goal was achieved in Chapter 5. The goal was achieved by designing an evaluation approach, running a large-scale reading comprehension experiment, and analysing the data obtained with two different evaluation metrics. Due to the fact that the majority of the participants were highly competent readers, the results showed no clear indication of an impact of the simplification approach on reading comprehension for all the participants as single group. Nevertheless, the results showed clear positive impact of the simplification approach on the reading performance of specific groups of readers.

Goal 7 was to evaluate the impact of the applied approach for writing and simplifying texts on other tasks which are important for the domain. The goal was achieved in Chapter 6. The tasks selected were manual translation and machine translation, due to the facts that in the modern global world emergency documents need to be translated and that these kinds of translation are most available to the general public. The evaluation was conducted by running an experiment with twenty-five professional translators and the machine translation engine Google Translate. The results for manual translation showed positive, but not satisfactorily statistically significant, impact of the simplification approach. The results for machine translation showed statistically significant positive impact of the simplification approach on this task.

Goal 8 was to evaluate the acceptability of the proposed approach for writing and simplifying texts with end-users and to identify its concrete weaknesses in terms of applicability. The goal was achieved in Chapter 7. It was achieved by running an experiment with six professional linguists. They were asked to read the controlled language guidelines and simplify documents according to them. The results produced a ranking of the controlled language rules in terms of how difficult they are to apply manually, as well as additional suggestions for future implementation priorities.

Goal 9 was to identify any weaknesses and limitations of the methodologies proposed in Goals 4-8 and to identify directions for improvement and future research. This goal was achieved in each of Chapters 3-7 by analysing in detail the shortcomings of the proposed approaches. On the basis of this analysis, a list of directions for future work was produced. It will be presented in this chapter.

8.2. Original Contributions of the Thesis

Achieving the goals described in Section 8.1 led to several original contributions affecting the NLP field in general, several NLP sub-fields, and other scientific fields. The **main contributions** are presented in the order of their appearance in the chapters of the thesis.

Contribution 1: The first corpus of crisis management documents in English (Crisis Management Corpus, CMC).

This new linguistic resource fills the gap in linguistic resources for NLP for Crisis Management and can be used both for developing NLP applications for the crisis management domain and for linguistic studies focussed on the sub-language of the crisis management field. The corpus has the size of 2 728 540 words and is composed of four sub-corpora, aiming to represent the variety of

crisis management document target audiences (general population/specialists), sub-domains (general crisis management/aeronautics/medical) and document types (instructions/alerts). Specifically, the first sub-corpus contains emergency instructions for the general population, the second sub-corpus contains instructions and protocols for crisis managers, the third sub-corpus contains instructions for pilots, and the fourth sub-corpus is composed of medical alerts. The corpus is in machine-processable format and has been pre-processed using the dependency parser *Machinese Syntax* (Tapanainen and Järvinen, 1997).

Contribution 2: The first numerical text complexity analysis of documents of the Crisis Management domain.

This analysis was made possible by the collection of the above-mentioned Crisis Management Corpus. It fills a gap in research on the communication efficiency of crisis management documents written in English. The analysis was focussed on two sets of surface linguistic markers—primary and secondary—based on existing literature about factors affecting reading comprehension. It aimed to test the research hypothesis that crisis management documents are too complex and need simplification. The testing was achieved by comparing the number of the two sets of linguistic markers in the CMC and Simple English Wikipedia. The results, statistically significant with 99% confidence, showed that the crisis management documents indeed exhibit a large number of high text complexity issues and need simplification. It also provided a clear picture of the text complexity issues affecting different types of crisis management texts. This contribution is useful for the NLP field and general Linguistics.

Contribution 3: Sub-language analysis of documents of the Crisis Management domain.

This analysis aimed to test the research hypothesis that crisis management documents exhibit linguistic features differing from those of general English. The analysis also employed the Crisis Management Corpus, but compared it with a corpus of general English—a random sample of the British National Corpus (BNC). The analysis was focussed on the number of the same two sets of high text complexity features as in the previous analysis, plus additional purely linguistic features, such as proportions of verbs, adjectives, adverbs, and nouns. The results, which are significant with 99% confidence, confirmed the research hypothesis. The analysis also showed interesting numerical differences between the combinations of linguistic features characterizing the separate CMC sub-corpora. This analysis contributes to NLP applications focussing on the crisis management field and to linguistic studies of the crisis management sub-language.

Contribution 4: An adaptation to English of a text simplification and document writing approach for crisis management documents.

The new approach is based on an existing controlled language for French, developed in collaboration with crisis management specialists. The resulting controlled language (Controlled Language for Crisis Management, CLCM) was developed specifically for the document type Instructions for the General Population and allows rewriting existing complex documents into clear ones and producing new clear crisis management documents in English. The importance of this approach is that it is tailored to the situational circumstances of reading and to the linguistic and textual characteristics of documents in the domain. CLCM is described in thirty-five pages of guidelines, containing (re-)writing rules, examples, lists of allowed and forbidden syntactic structures and lexical expressions, a grammatical term dictionary, and a domain dictionary. It is

argued that the controlled language should be easily adaptable to other types of documents from the domain and to similar types of documents from other domains. This contribution bears importance for the linguistic, NLP and psycholinguistic fields, as well as to the natural language generation sub-field of NLP.

Contribution 5: Transfer of the CLCM rules from English to Bulgarian.

This transfer resulted in a draft of the first controlled language addressing the complexity and ambiguity of emergency instructions in Bulgarian. This is a significant contribution, because emergency instructions in Bulgarian have never been treated before, and they still reflect the language which was typical more than twenty years ago. The results from this adaptation were presented to Bulgarian crisis management specialists. The results encountered high interest and received positive feedback. This transfer is a contribution to the Bulgarian linguistics and NLP fields.

Contribution 6: A new evaluation perspective of controlled language guidelines

Due to the fact that inaccuracies of communication in the crisis management domain can lead to loss of lives, an extensive evaluation of CLCM from multiple perspectives was required. The evaluation approach was based on the evaluation methodology proposed by Hirshman and Mani (2001) for assessing the output of NLP systems—specifically, the CLCM simplification was evaluated on extrinsic tasks (text complexity, reading comprehension, manual translation, and machine translation). In addition, the method of CLCM simplification was evaluated in terms of users' acceptability. The evaluation employed existing and developed new techniques. Such an extensive evaluation of a text simplification or controlled language approach has never been

conducted before. This is a contribution to the fields of natural language generation, text simplification and controlled languages evaluation.

Contribution 7: An adaptation of a standard psycholinguistic method to measuring the impact of text simplification approaches on reading comprehension under stress.

This method was developed with the aim of testing the impact of CLCM on the task which is its main purpose: the enhancement of reading comprehension in emergency situations. The evaluation method is based on reading texts and answering questions about them afterwards. A simulation of a stress situation is ensured by limiting the time allowed for reading the texts. Reading comprehension is measured via two metrics—percentage of correct answers and time employed to provide correct answers. The results of the evaluation showed clear improvement of reading comprehension of the simplified text for some groups of readers. This contribution is important for any new text simplification, controlled language, or natural language generation approaches addressing reading in emergency or normal situations, as well as for psycholinguistic studies analysing the impact of text simplification on reading comprehension.

Contribution 8: Novel findings about the general and CLCM-influenced reading comprehension of specific groups of readers.

These findings were obtained during the evaluation of the impact of CLCM simplification on reading comprehension. They provide interesting information about the differences in reading of groups of readers divided by human variables (age, sex, profession, and native language). In particular, it has been shown that female participants reply to questions faster than males, but are not affected by text simplification, while male participants reply slower to questions after reading

complex text, and their time of replying diminishes after reading simplified text. The results also showed significantly increased time to reply for subjects with ages over 45 years old, and very clearly revealed a positive impact on the percentage of correctly answered questions for groups of readers such as computational linguists, translators, linguists, and lawyers, as well as native speakers of non-Indo-European languages. These findings bear important contributions to the psycholinguistics, sociolinguistics, and educational fields.

Contribution 9: The finding that text simplification increases the difference in speed between post-editing machine translation (MT) and manual translation.

The comparison of the times taken to manually translate texts and to manually post-edit the MT output of the same texts confirmed the finding of Sousa et al. (2011) that post-editing machine translation is faster than translating from scratch. The results show that although for both complex and simplified texts post-editing is faster, text simplification makes the improvement larger for the simplified text (complex text post-editing 17.2% faster, simplified text post-editing 34.8% faster). This finding, which is important for translation technologies, is additionally enriched by the results obtained in the context of comparing a complex and a simplified text.

Contribution 10: An innovative cognitive evaluation approach for machine translation (MT) post-editing.

Due to the fact that the temporal and technical evaluations of the impact of CLCM on MT output provided positive but different results for the data on different languages, there was a need to analyse separately and concretely the post-edited versions of the different languages. The third evaluation approach is an innovative cognitive evaluation approach which is based on the

assumption that if CLCM has a positive impact on MT engine performance, then the cognitive effort required for the human post-editors to post-edit the translation of the simplified text will be less than the cognitive effort required for post-editing the translation of the complex text. The cognitive effort is calculated on the basis of the number of easy- and difficult-to-correct errors in the MT output based on complex and simplified texts. The approach provides an alternative to the shortcomings of the existing MT post-editing cognitive evaluation approaches and can also be used for evaluation of machine translation. The results of the evaluation showed that less cognitive effort is needed for post-editing MT output based on simplified text.

Contribution 11: A new numerical approach for evaluation of controlled language acceptability to users.

To evaluate the CLCM guidelines and rules, a detailed study of the internal process of manual simplification was conducted. The results revealed interesting findings about the time employed to simplify texts, the differences between different simplified versions of the same complex text, and the most difficult rules to apply for manually simplifying texts. The results revealed that the CLCM rules which are the most difficult to apply involve rewriting the text complexity issues which are the harder from the cognitive point of view. Such an internal view of the process of text simplification according to controlled language guidelines has never been conducted before.

In summary, this thesis brings contributions of different kinds (evaluation approaches, findings, studies, and resources) to several scientific fields, including **NLP, Linguistics, Psycholinguistics,** S**ociolinguistics**, and **Education**, as well as to the **Text Simplification**, **Machine Translation, Natural Language Generation** and **Translation Technologies** sub-fields of NLP. The next section will provide a review of the thesis.

8.3. Review of the Thesis

This section provides a brief review of the first seven chapters of this thesis.

Chapter 1 presented the context of and motivations for the research presented in this thesis. The main aims of the thesis were introduced, as well as the research hypotheses on which the research is based. A list of goals to be achieved in order to fulfil the aims of the thesis and the contributions which they would generate followed. Finally, the chapter introduced the structure of the thesis and the contents of each chapter.

Chapter 2 presented the problem addressed by this thesis (Text Complexity, TC) and its proposed solution (Text Simplification, TS). Then it provided an overview of the earlier and modern approaches to measuring text complexity and a detailed discussion of the existing text simplification approaches. The TS approaches were classified on the basis of the degree of their automation and with respect to whether they involved controlled languages or not. The limitations of the existing approaches with respect to application for the purposes of this thesis were explicated.

Chapter 3 introduced the Crisis Management Corpus and its text complexity and linguistic analyses, motivated by the demonstration in the previous chapter of the inadequacy of the existing approaches for measuring TC. The methods of collection and pre-processing of the corpus were described, along its composition and distribution of documents per sub-corpus. The two corpus analyses were based on two research hypotheses and employed a set of TC and linguistic features whose presence in the texts was analysed. The analyses employed Python scripts developed by the author. Analysis of the results showed that crisis management documents exhibit high levels of text complexity and different linguistic features than general English and thus need a specific text

simplification approach or guidelines for writing clear documents.

Chapter 4 described the Controlled Language for Crisis Management, which is a text simplification and documents writing approach tailored for the crisis management domain and adapted from an existing controlled language for French. Since the analysis in the previous chapter showed that the sub-corpus *Instructions for the General Population*, which represents the weakest link in crisis management communication, exhibits a higher number of high TC issues than the others, CLCM addressed this document type. A small experiment aiming to transfer the CLCM rules to an under-resourced language was mentioned. Finally, formal comparison of CLCM with other controlled languages, including the controlled language from which it was adapted, was performed, opening the way to further CLCM evaluations. The need for further evaluations was motivated by the crucial role of emergency instructions in preserving people's lives.

Chapter 5 performed the evaluation of CLCM on its most important task: reading comprehension under stress. A standard psycholinguistic approach was adapted to the needs of the thesis. The *"Online Reading Comprehension Experiment"* involving 103 participants with high reading skills showed that although there is no clear impact of the controlled language simplification on reading comprehension for the group as a whole, specific groups of readers do benefit. The metrics used for the analysis (percentage of correct answers and time to reply correctly) were ultimately combined into a novel unified evaluation metric, called C-factor.

Chapter 6 conducted the evaluation of the CLCM simplification on another task which is important for the domain—translation. Due to reasons of availability to the general public, which is the intended audience to which emergency instructions are addressed, the two types of translation investigated were manual and machine translation. The analysis applied one evaluation metric for

manual translation (time) and three metrics for machine translation (time, amount of post-editing, and types of corrected errors). The results showed a decisive positive impact of CLCM on machine translation and confirmed a prior finding that machine translation is faster than manual translation.

Chapter 7 presented the final evaluation of the CLCM simplification. After the positive results of the previous two evaluations, the need for a concrete analysis of the internal process of simplification was seen. The evaluation consisted of the *"Text Simplification Task Experiment"* with six natural language processing specialists with a background in Linguistics, their simplifications of multiple complex texts, and feedback on a questionnaire. The evaluation lead to findings regarding the time necessary to simplify instructions, the rules which are the most difficult to apply, and suggestions for future implementation priorities which would speed up the simplification process and make it uniform.

8.4. Directions for Future Work

This section provides and discusses a list of possible directions for future work. They either emerged during the study or are motivated by weaknesses of the proposed methodology. The issues are grouped according to the Chapter to which they are related.

- **Chapter 3: Further study of the crisis management sub-language and the crisis management corpus text complexity analysis**

  1. **Study of the crisis management sub-language**

As explained in Section 1.1, although the crisis management field is developing very quickly, there

is still a low number of linguistic and NLP applications developed for this field. For this reason, future work on the crisis management language in terms of terminology, discourse, syntax is essential. This would assist the developing on future NLP applications tailored for this field.

2. Further work on the relations between contradictory high text complexity issues

As explained in Section 3.3.3.1, some factors which are considered to be high text complexity issues have contradictory relationships. An example is “word length versus number of word senses”. The contradiction is that although short words are considered to be more comprehensible than long ones, it can be inferred from the Zipf's law (Zipf, 1949) that shorter words have higher frequency and that they also usually have a higher number of word senses. Future work could include an investigation of the relationship between these two factors and ways to overcome this issue, since it is important for the precise estimation of the text complexity of a given text.

3. **Further work on the methodology for detecting high text complexity markers**

As described in Section 3.3.3.2, the current method for detecting some of the secondary high TC issues (subordination markers, relative clause markers, and discourse markers) is coarse-grained and does not disambiguate their correct attribution to the respective category sufficiently precisely. As explained in Section 3.4, future work may include disambiguating them by building a complex grammar or using additional resources. As is the case with the previous issue, this is important for the precise estimation of the text complexity level of a given text.

- **Chapter 4: Future work on the Controlled Language for Crisis Management**

1. Continuing the development of the controlled language for Bulgarian

As explained in Section 4.3.4, some work has been done on adapting CLCM to Bulgarian language in which the emergency instructions are still written in a language typical for over twenty years ago. As the whole Bulgarian crisis management infrastructure is currently under development, future work will include continuing the work and completing the guidelines of the controlled language for Bulgarian in collaboration with the Bulgarian Academy of Sciences and the Bulgarian Crisis Management stakeholders. Also, now that evaluation methodologies have been established by this thesis, future work will include testing the quality of the controlled language for Bulgarian.

2. Testing CLCM portability

As mentioned in Section 4.5, CLCM should be easy to adapt to other documents from the crisis management domain or to documents of a similar type (i.e. instructions). This is due to the fact that although it is focussed on instructions for the general population, it addresses several high text complexity issues which can also affect other types of documents, such as long words, long sentences, ambiguous words, vague quantifiers, inconsistent terminology, complicated syntax, passive voice, negative constructions, illogical order of statements, unclear anaphoric links, and missing discourse connectives. Future work could include testing the controlled language with respect to its portability to other types of documents, such as medical leaflets.

- **Chapter 5: Improving the methodology of evaluation in the reading comprehension experiment**

1. Further diversifying the participants in the experiment

As seen in Section 5.5.2.1, due to the method of advertising the experiment**,** most of the participants in the experiment have the same high level of reading skills, and this is a likely cause of a lack of effect of CLCM on the full group of participants. Since specific groups of participants showed a positive impact of CLCM, in order to have a more representative picture of the impact of CLCM on reading comprehension, a more diversified and stratified sample of participants is needed. Future work should include social groups with lower literacy levels, lower reading skills, and less general knowledge, in order to test more thoroughly whether the proposed text simplification re-writing method is successful.

2. Taking into account the longer visual length of the simplified texts

As was explained in Section 5.5.2.1, and noted by some participants, the simplified texts have much greater visual length on the page than the complex ones, and participants need to scroll down the page, leading to insufficient time for reading some of the texts. In order to ensure more precise evaluation, future work could include this as a penalty for the simplified text.

3. Future analysis of specific sets of texts

As was seen from the results presented in Section 5.4, the CLCM simplification had a negative

impact on Set 2 and a substantially positive impact on Set 4. Future work may include further textual examination to find aspects of the input texts which contribute to why these two sets behave differently.

4. Taking into account the length of questions and answers

Due to the fact that different questions and answers had substantially different lengths, and the fact that the time to reply to questions includes the time for reading them at least once, future work could include taking these lengths into account when evaluating the time, in order to obtain a more precise evaluation. Calculating the text complexity levels of questions and answers is an additional idea for future research.

5. Implementing the C-factor

As explained in Section 5.5.2.2, since the evaluation of reading comprehension took into account two criteria—the percentage of correct answers and the time to reply correctly to questions—an idea for future work would be to combine both measures into one and obtain a unique reading comprehension factor (called C-factor), which would depend directly on the proportion of correct answers and inversely on the mean time to provide correct answers. Ranking comprehension of texts according to it for different groups of readers may produce interesting results.

- **Chapter 6: Further work related to the machine translation evaluation**

1. Evaluation of the impact of CLCM on translation of larger data sets

As seen in Chapter 6, the experiment was conducted on two texts of 150 words each and involved

only between three and five translators per language, which caused low statistical power. For this reason, it is necessary to conduct future evaluations on larger texts, including a higher number of translators and applying the cognitive evaluation method on more languages.

2. Allowing the use of more translation resources for proper manual evaluation

As explained in Section 6.5.4, translators employed no additional resources, e.g. dictionaries or terminological databases, to assist them with translation. Future work could include involving such resources or even translation memories to test whether the results for manual translation improve in this way.

3. Improving the interface for carrying out the experiment in Chapter 6

As explained in Section 6.5.4, the participants found the interface for this experiment difficult to use. Any future work could include upgrading to a more user-friendly interface, such as the one used in Aziz et al. (2012).

4. Taking into account the language transfer between pairs of languages

The results for manual translation gave very different results for different target languages, despite the fact that we assume that the translators had very similar skills. Therefore, in addition to testing with a larger number of participants in order to avoid the disproportional effect of outliers, a further study of the difficulty of translating standard language for the specific language pairs is necessary. Future work may also include evaluating the impact of CLCM on the translation process of each specific language pair.

5. **Taking into account the quality of Google Translate translation pairs**

As explained in Section 6.5.4, an evaluation of the Google Translate language pairs could provide explanations of the variability of machine translation results for different languages.

- **Chapter 7: Improvement of the evaluation experiment and future implementation**

1. **Use of advanced recording programs**

Since the relevant technology has advanced significantly, future work may include the use of software such as TransLog[44] to study the process of text simplification on computers using controlled language guidelines and eye-tracking technologies (similarly to Doherty et al., 2010).

2. **Provide better end-users training**

Future work could also include testing whether more appropriate and longer training of human simplifiers would change which rules are considered difficult to apply.

3. **Implementation of the CLCM writing aid**

As discussed in Section 7.7, future work can proceed with the implementation of the CLCM rules and assisting simplification operations identified by the results from Parts 2 and 3 of the questionnaire. For example, some candidates for implementation are the rules which are cognitively difficult to apply, such as replacing negative expressions with positive ones or highlighting ambiguous lexical terms.

44 http://www.translog.dk/. Last accessed on March 30th, 2012.

- **Further evaluation of CLCM**

Finally, future work could also include making a comparison of CLCM with other approaches to text simplification and text generation for lay readers. The evaluation methodologies that were developed in the course of this research, described in Chapters 5, 6, and 7, can be used for performing this comparison.

8.5. Thesis Final Remarks

Due to the multidisciplinary nature and the extensive scope of this research, future work, including improvements to the current methodology, can take a very high number of directions. Nevertheless, this thesis makes several significant contributions to communication management in emergency situations for such an international language as English. These contributions include the first text complexity analysis of the crisis management language, the first corpus of crisis management documents, the transfer of a controlled language from French to English, several evaluation methodologies which can be applied to other similar applications, and interesting psycholinguistic findings. With these results, methodologies, and resources, the thesis aims to make crisis management more efficient and to contribute to the safety and security of our modern world.

References

Aikawa, T. et al. 2007. "Impact of Controlled Language on Translation Quality and Post-Editing in a Statistical Machine Translation Environment". In Proceedings of the MT Summit XI. Copenhagen, Denmark, 10-14 September. 1-7.

Air Line Pilots Association. 1977. Aircraft accident report. Human factors report on Tenerife accident. Engineering and Air Safety, Washington D.C.

Al-Qiani, J. 2000. Translation Quality Assessment. Strategies, Parameters and Procedures. Meta: Translators' Journal, XLV, 3, 2000.

Allen, J. 2001. 'Post-Editing: An Integrated Part of a Translation Software Program'. Language International, April 2001, pp. 26-29.

Allen, J. 2002. Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes, book review, Multilingual Computing&Technology, 13.2, March 2002, 27-29.

Allen, J. 2003. Post-editing. In Computers and Translation: A Translators Guide. Edited by Harold Somers. Amsterdam: John Benjamins. chapter 16, pages 297-317.

Alqvist, I., and Sagvall Hein, A. 1996. Defining ScaniaSwedish - a Controlled Language for Truck Maintenance. Proceedings of the First International Workshop on Controlled Language

Applications CLAW96. Leuven, Belgium: Katholieke Universiteit Leuven Centre for Computational Linguistics, pp. 159-167.

Aluísio, S.M, Specia, L., Pardo, T.A.S., Maziero, E.G., Caseli, H.M., Fortes, R.P.M. 2008. A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps towards Text Simplification Systems. 26th ACM International Conference on Design of Communication SIGDOC-2008. pp. 15-22. Lisbon, Portugal.

Anderson, R. C., & Freebody, P. 1981. Vocabulary knowledge. Newark, DE: International Reading Association.

Andrews, S. 1997. The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. Psychonomic Bulletin and Review, 4, 439-461.

Angelov, K., Ranta, A., 2009. Implementing Controlled Languages in GF. Proceedings of CNL2009.

Antworth, E. L. & McConnel, S. R. 1992. KTEXT User's Guide. Summer Institute of Linguistics, Dallas.

Arnold, D., Balkan, L., Humphreys, R.L., Meijer, S. and Sadler, L. 1993. Machine Translation. An Introductory Guide. Blackwells-NCC, London, UK.

Aziz, W., Sousa, S. C. M., Specia, L. 2012. PET: a tool for post-editing and assessing machine translation. In The Eighth International Conference on Language Resources and Evaluation,

LREC ’12, Instambul, Turkey. May 2012.

Baddeley, A.D., & Hitch, G. 1974. Working memory. In G.H. Bower Ed..The psychology of learning and motivation: Advances in research and theory Vol. 8, pp. 47--89. New York: Academic Press.

Baldwin, B. 1995. CogNIAC: A Discourse Processing Engine. University of Pennsylvania.

Banko, M. and Brill, E. 2001. Scaling to very very large corpora for natural language disambiguation. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics ACL '01.Association for Computational Linguistics, Stroudsburg, PA, USA, 26-33.

Barry, C., Morrison, C.M. & Ellis, A.W. 1997. Naming the Snodgrass and Vanderwart pictures: Effects of age-of-acquisition, frequency and name agreement. Quarterly Journal of Experimental Psychology, 50A, 560-585.

Barthe K. et al. 1999. "GIFAS Rationalized French: A Controlled Language for Aerospace Documentation in French", Technical Communication, 46, 1999, pp. 220-229.

Barthe, K. 1996. EUROCASTLE – A User’s Experience with Prototype AECMA SE Checkers. Proceedings of the First International Workshop on Controlled Language Applications CLAW96.Leuven, Belgium: Katholieke Universiteit Leuven Centre for Computational Linguistics, pp. 42-63.

Barthe, K. 1998. GIFAS Rationalised French: Designing One Controlled Language to Match Another. Proceedings of the Second International Workshop on Controlled Language Applications CLAW98. Pittsburgh, Pennsylvania: Language Technologies Institute, Carnegie Mellon University, pp. 87-102.

Bateman, J. and Zock, M. 2003. Natural Language Generation. Oxford Handbook of Computational Linguistics. Chapter 15. Edited by R. Mitkov.

Bernth, A and Gdaniec, C. 2001. MTranslatability. Machine Translation. 2001, vol 16, number 3, pages 175 - 218. Kluwer Academic Publishers.

Bernth, A. 1999. 'A Confidence Index for Machine Translation'. In: Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation TMI 99. Chester, England, pp. 120–127.

Biran O., Brody S., and Elhadad, N. 2011. Putting it Simply: a Context-Aware Approach to Lexical Simplification. 2011. ACL, pp. 496-501. Portland, OR.

Bird, S., Klein, E. and Loper, E. 2009. Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit. O'Reilly Media.

Blanco X. 2009. Remarks about Linguistic Analysis, Normalization and Translation of Spanish "What to Do in Case of Fire" Texts, in ISMTCL Proceedings, International Review Bulag, PUFC, ISSN 0758 6787, ISBN 978-2-84867-261-8, pp 43-48

Bogacki K. 2009. Controlled Languages and Machine Translation, in ISMTCL Proceedings, International Review Bulag, PUFC, ISSN 0758 6787, ISBN 978-2-84867-261-8, pp 49-55

Bogacki K., Hebel K., 2009. Functional and Linguistic Characteristics of the ControllEdit Software.

Bogacki, K. 2009. “Vers une version contrôlée du polonais”, Panorama des etudes en linguistique diachronique et syntaxiques, Melanges offerts a Józef Sypnicki, edited by Grażyna Vetulani. Leksem, ask 2009, pp. 31-50.

Bowker, L. 2002. Computer-aided translation technology: A practical introduction. University of Ottawa Press, Ottawa, 2002.

Brandt S. 1997. Statistical and Computational Methods in Data Analysis, 3rd edn. Springer, New York

Briscoe, T. & Carroll, J. 1995. Developing and evaluating a probabilistic LR parser of part-of-speech and punctuation labels. In Proceedings of the 4th International Workshop on Parsing Technologies IWPT-95. pages 48--58, Prague/Karlovy Vary, Czech Republic.

Brown, G.D.A. & Watson, F.L. 1987. First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. Memory and Cognition, 15, 208-216.

Burnard, L. 1995. Users Reference Guide, British National Corpus Version 1.0. Oxford University Computing Service.

Burstein, J., Tetreault, J. and Andreyev, S. 2010. Using entity-based features to model coherence in student essays. In North American Chapter of the Association for Computational Linguistics, pages 681–684, Los Angeles, California. Computational Linguistics, 291.:19–51.

Bustamante, F. R., Declerck, T., and Leon, F. S. 2000. Towards a Theory of Textual Errors. Proceedings of the Third International Workshop on Controlled Language Applications CLAW00. Seattle, Washington: Association for Computational Linguistics, pp. 20-32.

Butler, C. S. 1985. Statistics in Linguistics. Oxford: Blackwell.

Canning, Y. 2002. Syntactic Simplification of Text. Ph.D. Theses, University of Sunderland.

Cardey S., 2011. Machine Translation of Controlled Languages for More Reliable Human Communication in Safety Critical Applications, in Proceedings of the 12th International Symposium on Social Communication - Comunicación Social en el Siglo XXI, Santiago de Cuba, Cuba, January 17-21, 2011, Vol. II, ISBN: 978-959-7174-19-6, pp. 953-958

Cardey, S., Bogacki, K., Blanco, X. and Mitkov, R. 2010. Resources for Controlled Languages for Alert Messages and Protocols in the European Perspective. In Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10. 19-21 May 2010, Valletta, Malta.

Carl M, Kay M, Jensen K. 2010. Long-distance Revisions in Drafting and Post-editing. Proceedings of CiCling 2010: 193-204.

Carroll, J. 2003. Oxford Handbook of Computational Linguistics. Edited by R. Mitkov. Chapter 12, Parsing.

Carroll, J. B. 1966. “An Experiment in Evaluating the Quality of Translations,” in Mechanical Translation 9, pp. 55-66.

Carroll, J.B. & White, M.N. 1973. Word frequency and age-of-acquisition as determiners of picture-naming latency. Quarterly Journal of Experimental Psychology, 25, 85-95.

Carver, R. P. 1992. Reading rate: Theory, research and practical implications. Journal of Reading, 36, 84-95.

Caseli, H.M., Pereira, T.F., Specia, L., Pardo, T.A.S., Gasperin, C., Aluísio, S.M. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In Alexander Gelbukh ed., Advances in Computational Linguistics, Research in Computer Science, vol 41, pp. 59-70. 10th Conference on Intelligent Text Processing and Computational Linguistics CICLing-2009., March 01–07, Mexico City.

Chandrasekar, R., Christine, D. and Srinivas, B. 1996. Motivations and methods for text simplification. In Proceedings of the16th conference on Computational linguistics – Volume 2, Edited by: Association for Computational Linguistics, Copenhagen, Denmark.

Chapman, W.W., et al. 2005. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artif Intell Med. 2005 Jan, 331.: 31-40.

Cholewa J., 2009. Remarks Concerning Writing Texts in Controlled Polish, in ISMTCL Proceedings, International Review Bulag, PUFC, ISSN 0758 6787, ISBN 978-2-84867-261-8, pp 62-68

Church, K. and Hovy, E. 1993. "Good Applications for Crummy Machine Translation". Machine Translation, 8 pp. 239–258.

Cohen, K.B., Johnson, H.L., Verspoor, K., Roeder, Ch. and Hunter, L.E. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. BMC Bioinformatics 11:492.

Coltheart, M. 1981. The MRC Psycholinguistic Database, Quarterly Journal of Experimental Psychology, 33A, 497-505.

Conway, M., Collier, N. and Doan, S. 2009. Using hedges to enhance a disease outbreak report text mining system. In BioNLP ’09: Proceedings of the Workshop on BioNLP, pages 142–143, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

Coppola, D.P. 2007. Introduction to international disaster management, Butterworth-Heinemann

Corston-Oliver, S. 2001. Text compaction for display on very small screens. In Proceedings of the workshop on automatic summarization NAACL'01. Pittsburgh, USA.

Corvey, W. J., Vieweg, S., Rood, T. and Palmer, M. 2010. Twitter in Mass Emergency: What NLP Techniques can Contribute. In Proceedings of the NAACL HLT 2010 Workshop on

Computational Linguistics in a World of Social Media Los Angeles, California, June 2010. 23–24.

Daelemans W., Höthker A. and Sang, E. T. K.2004. Automatic Sentence Simplification for Subtitling in Dutch and English. In Proceedings of the 4th International Conference on Language Resources and Evaluation, volume III, pages 1045-1048, Lisbon, Portugal, May 2004.

Dale, E. and Chall, J. 1948. A Formula for Predicting Readability. Educational Research Bulletin, 27, 28, pp.11-20, 37-54.

Dale, R., H. L. Somers, and Hermann Moisl Eds.). Marcel Dekker, Inc., New York, NY, USA.

Daume, H. I. and Marcu, D. 2005a. A Bayesian Model for Supervised Clustering with the Dirichlet Process Prior. Machine Learning Research, 6, pp.1551-1577.

Daume, H. I. and Marcu, D. 2005b. Induction of Word and Phrase Alignments for Automatic Document Summarization. Computational Linguistics, 314. pp.505-530.

Davis, L. E., LaTourrette, T., Mosher, D.E., Davis, L.M. and Howell, D.R. 2003. Individual Preparedness and Response to Chemical, Radiological, Nuclear, and Biological Terrorist Attacks. Santa Monica, CA: RAND Corporation, 2003.

Denzin, N. 1970. The research Act in Sociology. Butterworth, London.

Devlin, S. 1999. Automatic Language Simplification for Aphasic Readers. PhD. Theses, University of Sunderland.

Dobrovol'skij, Dmitrij O.Piirainen, Elisabeth 2005.: Figurative Language: Cross-cultural and Cross-linguistic Perspectives. Amsterdam [etc.]: Elsevier.

Doherty, S., O'Brien, S. and Carl, M. 2010. Eye tracking as an MT evaluation technique. Machine Translation, 24, 1, pp1-13.

Dorr, B., Zajic, D. and Schwartz R. 2003b. Hedge trimmer: A parse-and-trim approach to headline generation. In Proceedings of the HLT- NAACL 2003 Text Summarization Workshop, Edmonton, Alberta, Canada, pages 1–8.

Dras, M. 1999. Tree Adjoining Grammar and the Reluctant Paraphrasing of Text. PhD Theses, Macquarie University, Australia.

DuBay, W.H. 2004. The Principles of Readability. Costa Mesa, CA: Impact information.

Dunlavy, D. M., Conroy, J. M., Schlesinger, J. D., Goodman, S. A., Okurowski, M. E., O'Leary, D. P. and Halteren, H. V. 2003. Performance of a Three-Stage System for Multi-Document Summarization. In Proceedings of the Document Understanding Conference DUC 2003.

Elhadad, M. and Robin, J. 1992. Controlling content realization with functional unification grammar. Proceedings of the 6th International Workshop on Natural Language Generation: Aspects of Automated Natural Language Generation, p.89-104, April 05-07, 1992. Edited by:

Springer-Verlag.

Elhadad, N. 2006. User-Sensitive Text Summarization: Application to the Medical Domain. Ph.D. Thesis, Columbia University, January 2006.

Euler, T. 2002. Tailoring Text Using Topic Words: Selection and Compression. DEXA '02: Proceedings of the 13th International Workshop on Database and Expert SystemsApplications: IEEE Computer Society, pp. 215-222.

Fellbaum, C. 1998. WordNet: an Electronic Lexical Database, MIT Press.

Flesch, R. 1948. A new readability yardstick. Journal of Applied Psychology, 32, pp.221-233.

Forster, K.I. & Chambers, S.M. 1973. Lexical access and naming time. Journal of Verbal Learning and Verbal Behaviour, 12, 627-635.

Fromkin, V. and Rodman, R. 1978. An introduction to language, 2nd edition. Holt, Rinehart and Winston.

Gasperin, C., Maziero, E., Specia, L., Pardo, T.S.P., Aluisio, S.M. 2009b. Natural language processing for social inclusion: a text simplification architecture for different literacy levels. XXXVI Seminário Integrado de Software e Hardware SEMISH-2009. pp. 387-401. Bento Gonçalves, Brazil.

Gasperin, C., Specia, L., Pereira, T., Aluisio, S.M. 2009a. Learning When to Simplify Sentences for

Natural Text Simplification. Encontro Nacional de Inteligência Artificial ENIA-2009. pp. 809-818. Bento Gonçalves, Brazil.

Gdaniec, C. 1994., 'The Logos Translatability Index'. In: Technology Paratnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in The Americas. Columbia, Maryland, pp. 97–105.

Gerber, L. and Hovy, E. H. 1998. Improving Translation Quality by Manipulating Sentence Length. Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup: Springer-Verlag, pp.448 – 460.

Gerrig R. 1986. Processes and products of lexical access. Language and Cognitive Processes. 1, 187-196.

Glenberg, A. M., Robertson, D. A., Jansen, J. L., & Johnson-Glenberg, M. C.1999. Not propositions. Journal of Cognitive Systems Research, 1, 19-33.

Goyvaerts P. 1996. Controlled English, curse or blessing? A user's perspective. Proceedings of CLAW 1996.

Graesser, A.C., Cai, Z., Louwerse, M., Daniel, F. 2006. Question Understanding Aid QUAID.: A web facility that helps survey methodologists improve the comprehensibility of questions. Public Opinion Quarterly, 70, 1-20.

Grefenstette, G. 1998 of Conference. Producing Intelligent Telegraphic Text Reduction to provide

an Audio Scanning Service for the Blind. In Working Notes of the AIII Spring Symposium on Intelligent Text Summarization, Stanford University, CA.

Griffith, A. 2007. The Risks of Using Spreadsheets for Statistical Analysis. IBM. SPSS Statistics.

Gronlund, N.E. 1982.Constructing Achievement Tests. Prentice-Hall, Englewood Cliffs, N.J. Third Edtiion.

Guerberof, Ana. 2009. Productivity and quality in the post-editing of outputs from translation memories and machine translation. Localisation Focus. The International Journal of Localisation. Vol. 7 Issue 1.

Gunning, R. 1952. The technique of clear writing. New-York: McGraw-Hill.

Gusfield, D. 1997. Algorithms on strings, trees, and sequences: computer science and computational biology. Cambridge University Press, New York, NY, USA.

Gwiazdecka E., 2009. Annotation of Terminology from Protocols in Polish Controlled Language, in ISMTCL Proceedings, International Review Bulag, PUFC, ISSN 0758 6787, ISBN 978-2-84867-261-8, pp 121-126

Hale, S. and Campbell, S. 2002. The Interaction Between Text Difficulty and Translation Accuracy. Babel, Volume 48, Number 1, 2002 , pp. 14-3320. John Benjamins Publishing Company.

Harley, T.A. 2008.The Psychology of the Language: from data to theory, Psychology Press, Hove

and New York, 2008.

Hatim, B., and Munday, J. 2004. Translation. An advanced resource book. London. Routledge.

Heurley L. 2001. Compréhension et utilisation de textes procéduraux : l'effet de l'ordre de mention des informations, Revue Française de Linguistique Appliquée 2001/2, Volume VI, pp. 29-46.

Hirschman, L., and Mani, I. 2001. Evaluation. In R. Mitkov, ed., Handbook of Computational Linguistics, Oxford University Press, 2003.

House, J. 2001. "Translation Quality Assessment: Linguistic Description versus Social Evaluation". Meta: Translators' Journal, vol. 46, n° 2, 2001, p. 243-257.

Howes, D.H. & Solomon, R.L. 1951. Visual duration threshold as a function of word probability. Journal of Experimental Psychology, 41, 401-410.

Huijsen, W.O. 1998. Controlled Language – An Introduction. Proceedings of the Second International Workshop on Controlled Language Applications. Piitsburgh, Pennsylvania.

Hutchins. J.W. & Somers, H.L. 1992. An Introduction to Machine Translation, London: Academic Press.

Ilisei, I., Inkpen, D., Pastor, G., and Mitkov, R. 2009. Towards Simplification: A Supervised Learning Approach, in Proceedings of Machine Translation Twenty-Five Years On, London, United Kingdom, November 21-22, 2009.

Inui, K., Fujita, A., Takahashi, T., Iida, R. and Iwakura, T. 2003 . Text simplification for reading assistance: A project note. Second International Workshop on Paraphrasing.

Ireson, N. 2009.Local Community Situation Awareness During an Emergency. Proceedings of the IEEE International Conference on Digital Ecosystems and Technologies IEEE-DEST 2009.

James C.T. 1975. The role of semantic information in lexical decisions. J Exp Psychol: Hum Percept Perform 104: 130–6.

Jang, H., Lim, J., Lim, J.-H., Park, S.-J., Lee, K.-C. and Park, S.-H. 2006. Finding the evidence for protein-protein interactions from PubMed abstracts. Bioinformatics, 2214. pp.220-226.

Johnson E. 1993. Talking across Frontiers. International Conference on European Cross Border Cooperation: Lessons for and from Ireland. Queen's University Belfast.

Johnson-Laird, P. N. 1975. Meaning and the mental lexicon. In A. Kennedy and A. Wilkes Eds.. Studies in long-term memory. pp. 123-142. London: John Wiley.

Johnson, E., M. Garner, S. Hick, and D. Matthews. 1993. 'PoliceSpeak - Police Communications and Language and the Channel Tunnel - Research Report', PoliceSpeak Publications, Cambridge.

Jurafsky, D. and Martin, J. H. 2008. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd ed., Prentice-Hall.

Kienzle, J., Guelfi, N., Mustafiz, S. 2010. Crisis management systems: A case study for aspect-oriented modeling. Transactions on Aspect-Oriented Software Development 7 2010. 1–22.

Kincaid, J. P., Fishburne, R. P. J., Rogers, R. L. and Chissom, B. S. 1975. Derivation of new readability formulas Automated Readability Index, Fog Count and Flesch Reading Ease Formula. for Navy enlisted personnel. Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

Kittredge, R. I. 2003. Oxford Handbook of Computational Linguistics. Edited by R. Mitkov. Chapter 23, Sub-languages and controlled languages.

Kiwan, D., Ahmed, A. and Pollitt, A. 1999. The effects of text comprehension and performance in examinations. Proceedings of BPS London Conference, December, 1999.

Klare, G. R., Sinaiko, H. W. and Stolurow, L. M. 1972. The cloze procedure: a convenient readability test for training materials and translations. Applied Psychology, 21: 77–105.

Klebanov, B., Knight, K. and Marcu, D. 2004. Text Simplification for Information Seeking Applications. On the Move to Meaningful Internet Systems, Lecture Notes in Computer Science, 3290, pp.735-747.

Knight, K. and Marcu, D. 2002. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. Artificial Intelligence, 1391. pp.91-107.

Krings, H.P. 2001. Repairing Texts: Empirical Investigations of Machine Translation Post-Editing

Processes. Koby, G.S. ed.. Kent, Ohio: The Kent State University.

Kuhn, T. 2007. AceRules: Executing Rules in Controlled Natural Language. Proc. First International Conference on Web Reasoning and Rule Systems RR2007. Springer, 2007.

Kuhn, T. 2009a. Controlled English for Knowledge Representation. Ph.D. Thesis. University of Zurich, Switzerland.

Kuhn, T. 2009b. How controlled English can improve semantic wikis. In proceedings of SemWiki2009: The Fourth Workshop on Semantic Wikis.

Kuhn, T. 2010. An Evaluation Framework for Controlled Natural Languages. In Proceedings of the Workshop on Controlled Natural Language CNL 2009.Springer, 2010.

LaFontaine, D., Monseur, C. 2009.Gender Gap in Comparative Studies of Reading Comprehension: to what extent do the test characteristics make a difference?, European Educational Research Journal, 81., 69-79. http://dx.doi.org/10.2304/eerj.2009.8.1.69

Language Processing, Speech Recognition, and Computational Linguistics. Chapter 20. 1st ed.,

Larigauderie P, Gaonac'h D, Lacroix N. Working memory and error detection in texts: what are the roles of the central executive and the phonological loop? Applied Cognitive Psychology, 1998, 12: 505~527.

Leroy G., Helmreich S., and Cowie J.R. 2010. "The Effects of Linguistic Features and Evaluation

Perspective on Perceived Difficulty of Medical Text," Hawaii International Conference on System Sciences, January 5-8, Kauai, 2010.

Leroy, G. 2010. Old and New Readability Metrics and their Relation to Text Difficulty", IHA's Ninth Annual Health Literacy Conference - Health Literacy in the Real World: Programs & Solutions That Work, Irvine, California, May 6-7, 2010.

Liben-Nowell, D. 2000.Syntactic Simplification. MSc. Thesis. University of Cambridge, UK.

Lietz, P. 2006. A meta-analysis of gender differences in reading achievement at the secondary school level. Studies In Educational Evaluation, Volume 32, Issue 4, 2006, Pages 317-344, ISSN 0191-491X.

Lönneker-Rodman, B. 2007. Advanced course Figurative Language Processing, 19th European Summer School in Logic, Language and Information ESSLLI 2007. Trinity College, Dublin, Ireland. 6-10 August 2007.

Lönneker-Rodman, B. and Mohit, B. 2008.: "Translation of the non-literal: Evidence from aligned bilingual corpora." In Abstract proceedings of the third international conference of the German Cognitive Linguistics Association DGKL/GCLA-2008. Leipzig, Germany, September 25-27, 2008. 156-157.

Lorge, I. 1948. The Lorge and Flesch Readability Formulae: A Correction. School and Society, Vol. 67, pp. 141-142.

Lux, V. and Dauphin, E. 1996. Corpora studies: a contribution to the definition of a controlled language. In Proceedings of the First International Workshop on Controlled Language Applications, pp 193204. Leuven, Belgium, March 1996.

Madoff L.C. 2004. ProMED-mail: an early warning system for emerging diseases. Clin Infect Dis, 2004,392.:227-32.

Mark, G., Bagdouri, M., Palen, L., Martin, J.H., Al-Ani, B., Anderson, K. to appear 2012). Blogs as a Collective War Diary. To appear in 2012 ACM Conference on Computer Supported Cooperative Work, Bellevue, WA.

Marslen-Wilson,W. 1990. Activation, competition, and frequency in lexical access. In G. Altmann Ed.. Cognitive Models of speech processing, pp. 148-172. Cambridge: MIT Press.

Martin, G. L. 2004. Encoder: A connectionist model of how learning to visually encode fixated text images improves reading fluency. Psychological Review, 111, 617-639.

Mason, J. M. and Kendall, J. R. 1978. Facilitating Reading Comprehension through Text Structure Manipulation. Technical Report No. 92. Champaign, Ill. Center for the Study of Reading, University of Illinois.

Mateescu A. and Salomaa A. 1997. Formal languages: an introduction and a synopsis. In G. Rozenberg and A. Salomaa, editors, Handbook of Formal Languages - Volume 1: Word, Language, Grammar, pages 1–40. Springer.

Max, A. 2005. Simplification interactive pour la production de textes adaptés aux personnes souffrant de troubles de la compréhension, Proceedings of TALN'05, Dourdan, France.

McDonald, D.D. 2000. Natural language generation. Handbook of Natural Language Processing.

McEnery, T. and Wilson, A. 1996. Corpus Linguistics. Edinburgh: Edinburgh University Press.

Mclaughlin, G. H. 1969. SMOG Grading — a New Readability Formula. Journal of Reading 128. pp.639-646.

McNamara, D.S., Louwerse, M.M., McCarthy, P.M., & Graesser, A.C. 2010. Coh-Metrix: Capturing linguistic features of cohesion. Discourse Processes, 47, 292-330.

Miller, G. A. 1956. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. The Psychological Review, 63, pp.81-97.

Mitamura, T. & Nyberg, E. 2001. Automatic rewriting for controlled language translation. Proceedings of the NLPRS-2001 Workshop on Automatic Paraphrasing: Theories and Applications. Tokyo, Japan. Pages 1-12.

Mitkov, R. 2002.Anaphora Resolution. Longman.

Mitkov, R. editor.The Oxford Handbook of Computational Linguistics Oxford Handbooks in Linguistics S.. Oxford University Press, 2003

Muegge, Uwe 2006. "Fully automatic high quality machine translation of restricted text: A case study", Proceedings of the twenty-eighth international conference on translating and the computer, 16-17 November 2006, London: Aslib, pp. 16–17.

Murphy, D., 2000. Keeping translation technology under control. In: Machine Translation Review, issue 11: December 2000, pp. 11-13.

Nastase, V. and Szpakowicz, S. 2003. Augmenting WordNet's structure using LDOCE. In Proceedings of the 4th international conference on Computational linguistics and intelligent text processing CICLing'03. Alexander Gelbukh Ed.. Springer-Verlag, Berlin, Heidelberg, 281-294.

Nayak, S. 2011. Towards a Grounded Model for Ontological Metaphors. RANLP Student Research Workshop, Hisar, Bulgaria.

Nenkova, A., Chae, J., Louis, A. and Pitler, E. 2010. Structural Features for Predicting the Linguistic Quality of Text: Applications to Machine Translation, Automatic Summarization and Human-Authored Text In Emiel Krahmer and Mariet Theune, editors, Empirical Methods in Natural Language Generation: Data-oriented Methods and Empirical Evaluation, 2010.

Nicolosi, L., Harryman, E., Kresheck, J. 2003. Terminology of Communication Disorders: Speech, Language, Hearing. Lippincott Williams & Wilkins, Fifth edition October 20, 2003.

Nida, E. and Taber C. R. 1969. The Theory and Practice of Translation, Leiden: Brill, 218 p.

Norman S., Kemper S., Kynette D. 1992. Adults' reading comprehension: Effects of syntactic complexity and working memory. Journal of Gerontology: Psychological Sciences. 1992,47:P258–P265.

Nyberg E., Mitamura T. & Huijsen, W.O., 2003. Controlled language for authoring and translation. In: Harold Somers ed.. Computers and translation: a translator's guide Amsterdam/Philadelphia: John Benjamins Publishing Company, 2003., pp.245-281.

O'Brien, S. 2005. 'Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Text Translatability'. In Machine Translation. Vol. 19, No 1. pp. 37-58.

O'Brien, S. 2006. Controlled Language and Post-Editing. Multilingual, Issue 83, pp. 17-19.

Ogden, Ch. K. 1930. Basic English: a general introduction with rules and grammar, London,Kegan Paul, Trench, Trubner.

Ogrizek, M. and Guillery, J-M. 1999. Communicating in crisis.Transaction Publishers.

Paivio A. 1971. Imagery and verbal processes. New York: Holt, Rinehart & Winston.

Papadopoulou E., Puig Portella M., 2009. Abduction Alerts in Greek and Spanish, in ISMTCL Proceedings, International Review Bulag, PUFC, ISSN 0758 6787, ISBN 978-2-84867-261-8, pp 185-189.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. 2002. BLEU: a method for automatic evaluation

of machine translation in ACL-2002: 40th Annual meeting of the Association for Computational Linguistics pp. 311–318.

Pool, J. 2006. “Can Controlled Languages scale to the Web?”, CLAW 2006 at AMTA 5th International Workshop on Controlled Language Applications. Cambridge, Massachusetts, USA, 2006.

Poprat, M., Beisswanger, E., and Hahn, U. 2008. Building a BioWordNet by using WordNet's data formats and WordNet's software infrastructure: a failure story. In Software Engineering, Testing, and Quality Assurance for Natural Language Processing SETQA-NLP '08). Association for Computational Linguistics, Stroudsburg, PA, USA, 31-39.

Prentice-Hall.

Puig Portella M., Papadopoulou E. 2009. Treatment of the Imperative Forms in Automatic Translation between Catalan, Spanish and Greek, in ISMTCL Proceedings, International Review Bulag, PUFC, ISSN 0758 6787, ISBN 978-2-84867-261-8, pp 198-202

Quirk, R. 1985. A Comprehensive Grammar of the English Language. Longman.

Rabin, A. T. 1988. Determining Difficulty Levels of Text Written in Languages Other than English. In B. L. Zakaluk and Samuels, S. J. eds.. Readability: It's Past, Present, & Future. Newark, Delaware: International Reading Association.

Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research.

Psychological Bulletin, 124, 372-422.

Regester, M. & Larkin, J. 2005. Risk issues and crisis management. A casebook of best practice 3rd ed.). London: Kogan Page. National Research Council Committee on Risk Perception and Communication NRC. 1989). Improving Risk Communication. Washington, D.C. National Academic Press.

Reichle, E. D., Rayner, K., & Pollatsek, A. 2003. The E-Z Reader model of eye-movement control in reading: Comparisons to other models. Behavioral and Brain Sciences, 26, 44-526.

Renahy J. 2009. Controlled Languages: a Scientific Popularization through the Example of the Controlled Language ``LiSe''. ISMTCL Proceedings, International Review Bulag, pp 215-222.

Renahy J., et al. 2009. Controlled language norms for the redaction of security protocols: finding the median between system needs and user acceptability. 11th International Symposium on Social Communication, Santiago de Cuba, Cuba, January 19-23, 2009.

Renahy J., Thomas I., Chippeaux G., Germain B., Petiaux X., Rath B., De Grivel V., Cardey S., Vuitton, D.A. 2011. La « langue contrôlée » et l'informatisation de son utilisation au service de la qualité des textes médicaux et de la sécurité dans le domaine de la santé, communication aux Journées Francophones d'Informatique Médicale, Tunis, 31 mars, 1 et 2 avril 2011, à paraître in la collection «Informatique et Santé», Springer-Verlag.

Renahy J., Vuitton D.A., Rath B., Thomas I., De Grivel V., Cardey S., 2011. Communicating

vaccine safety: standardized language and automatic translations systems for safety protocols, à paraitre in Communicating Vaccine Safety, Rath B. et al., 2011

Renahy, J., Gin, J., Devitre, D., Beddar, M., Kiattibut-Ananta, R., Wu, W., Mikati, Z., De Grivel, V., Courtebras, V., Haeringer-Cholet, A., Aishan, A., Mahemuti, A., Sabbah, I., Chalermpantmetagul, S., Rath, B., Greenfield, P., Vuitton, D.A., Cardey, S. 2010. Development and Evaluation of a Controlled Language and of a computerized writing assistant "LiSe" to improve the quality and safety of medical protocols. International Forum on Quality and Safety of Health Care. 20-23 April 2010, The Nice Acropolis, Nice, France.

Robertson F.A., Johnson E. 1988. AirSpeak: Radiotelephony Communication for Pilots. Oxford, Prentice Hall

Roman, J.H., et al. 2008. Reducing information overload in emergencies by detecting themes in Web content. Proceedings of the 5th International ISCRAM Conference 2008,101-6.

Rudas Z., 2009. Polish Controlled Language and Its Machine Translation into French, in ISMTCL Proceedings, International Review Bulag, PUFC, ISSN 0758 6787, ISBN 978-2-84867-261-8, pp 231-235

Ruffino, J.R. 1982. Coping with machine translation. In: Lawson 1982. 57-60.

Sanz de Acedo Lizarraga, M.L., Sanz de Acedo Baquedano, M.T., Cardelle-Elawar M. 2007.Factors that affect decision making: gender and age differences. International Journal of Psychology and Psychological Therapy. 2007,73.:321–391.

Schactl, S. 1996. Requirements for Controlled German in Industrial Applications. Proceedings of the First International Workshop on Controlled Language Applications CLAW96. Leuven, Belgium: Katholieke Universiteit Leuven Centre for Computational Linguistics, pp. 143-149.

Schäfner, F. 2003. "MT post-editing: How to shed light on the "unknown task". Experiences made at SAP", Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, 15-17 May 2003, Dublin City University.

Schiaffino, R. and Zearo, F. 2006. Developing and using a translation quality index. MultiLingual magazine. July/August 2006 issue.

Schiffrin, D. 1987. Discourse Markers. Cambridge University Press.

Schneid, T.D. and Collins, L. 2001. Disaster management and preparedness. Lewis Publishers.

Schwitter, R. 2008. A Controlled Natural Language for the Semantic Web. Journal of Intelligent Systems, 17, pp.125-141.

Seeger, M. W., Sellnow, T. L., & Ulmer, R. R. 1998. Communication, organization and crisis. Communication Yearbook 21: 231–275.

Siddharthan, A. 2003. Syntactic Simplification and Text Cohesion. PhD. Theses. University of Cambridge, UK.

Siddharthan, A., Nenkova, A. and Mckeown, K. R. 2004. Syntactic Simplification For Improving Content Selection In Multi-Document Summarization. In Proceedings of the 20th International Conference on Computational Linguistics COLING 2004. pages 896-902, Geneva, Switzerland.

Snover, M., Madnani, N., Dorr B. and Schwartz, R. "Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric" 2009. In Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics EACL-2009).

Solheim, T., Lorentsen, M., Sundnes, P.K., Bang, G. & Bremnes, L. 1992. The “Scandinavian Star” ferry disaster 1990 - a challenge to forensic odontology. International Journal of Legal Medicine 104: 339-345.

Somers, H. and Wild, E. 2000.Evaluating Machine Translation: the Cloze procedure revisited. In Translating and the Computer 22: Proceedings of the Twenty-second International Conference on Translating and the Computer, London.

Sousa, S., Aziz, W., Specia, L. 2011. Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. Recent Advances in Natural Language Processing Conference RANLP-2011., September, Hissar, Bulgaria.

Sowa, J. F. 2004. Graphics and Languages for the Flexible Modular Framework. International Conference on Conceptual Structures ICCS. Edited by: Springer-Verlag Berlin.

Spache, G. 1953. A new readability formula for primary-grade reading materials. The Elementary School Journal 537. pp.410–413.

Specia, L. 2010.Translating from Complex to Simplified Sentences. 9th International Conference on Computational Processing of the Portuguese Language Propor-2010.Lecture Notes in Artificial Intelligence, Vol. 6001, Springer, pp. 30-39, Porto Alegre, Brazil

Steedman, M. 2008. On becoming a discipline. Computational Linguistics, 341.: 137–144, 2008.

Steinberger R., Pouliquen, B. & van der Goot, E. 2009. An introduction to the Europe Media Monitor Family of Applications. In: Fredric Gey, Noriko Kando & Jussi Karlgren eds..: Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop SIGIR CLIR'2009., pp. 1-8. Boston, USA. 23 July 2009.

Streiff, A.A. 1985. New developments in TITUS 4. In: Lawson 1985. 185-192.

Strevens, P. 1984. Seaspeak Reference Manual. Pergamon Press.

Szarvas, G., Vincze, V., Farkas, R., Csirik, J. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts, Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, June 19-19, 2008, Columbus, Ohio.

Szmrecsanyi, B. M. 2004. On operationalizing syntactic complexity. In: G. Purnelle, C. Fairon and A. Dister, Editors, Proceedings of the seventh international conference on textual data statistical analysis, Presses universitaires de Louvain. II, Louvain-la-Neuve 2004. pp. 1032–

1039.

Tapanainen P and Järvinen T. 1997. A non-projective dependency parser. In Proceedings of the 5th Conference on Applied Natural Language Processing: 31 March-3 April 1997, Washington D.C , pp. 64-71.

Tatsumi, M. 2010. Post-Editing Machine Translated Text in a Commercial Setting: Observation and Statistical Analysis. PhD thesis, Dublin City University.

Temnikova, I. 2010. A Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment. International Conference "Language Resources and Evaluation" LREC2010. Valletta, Malta. May 17-23, 2010.

Temnikova, I. 2011. Establishing Implementation Priorities in Aiding Writers of Controlled Crisis Management Texts. International Conference "Recent Advances in Natural Language Processing" RANLP 2011. Hissar, Bulgaria. September 12-14, 2011.

Temnikova, I. and Cohen K. B. 2012. The Crisis Management Corpus and its Application to the Study of the Crisis Management Sub-language. Accepted at the forthcoming workshop "Language Resources for Public Security Applications" at the International Conference "Language Resources and Evaluation" LREC 2012. Istanbul, Turkey. May 27, 2012.

Temnikova, I. and Margova, R. 2009. Towards a Controlled Language in Crisis Management: The Case of Bulgarian. Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages ISMTCL. Besancon, France, July 1-3, 2009.

Temnikova, I. and Orasan, C. 2009. Post-editing Experiments with MT for a Controlled Language. Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages ISMTCL. Besancon, France, July 1-3.

Temnikova, I. Orasan, C. and Mitkov, R. 2012. CLCM - A Linguistic Resource for Effective Simplification of Instructions in the Crisis Management Domain and its Evaluations. Accepted at the forthcoming International Conference "Language Resources and Evaluation" LREC 2012. Istanbul, Turkey. May 21-27, 2012.

Tweedie, F. J. and Baayen R. H. 1998. How Variable May a Constant Be? Measures of Lexical Richness in Perspective. Computers and the Humanities Vol. 32, No. 5 1998. pp. 323-352

Unwalla M. 2004. AECMA Simplified English. Journal Communicator, Winter 2004.

Van Oosten, P., Tanghe, D., & Hoste, V. 2010. Towards an Improved Methodology for Automated Readability Prediction. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias eds.. Proceedings of the seventh International Conference on Language Resources and Evaluation LREC'10. European Language Resources Association, Valletta, Malta.

Vander Linden, K. 2000. Generation. Speech and Language Processing: An Introduction to Natural

Vanderwende, L., Suzuki, H., Brockett, C. and Nenkova, A. 2007. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. Information Processing & Management, 436. pp.1606-1618.

Vassiliou, M., Markantonatou, S.,Maistros, Y., Karkaletsis, V. 2003. "Evaluating Specifications for Controlled Greek", EAMT-CLAW2003, Dublin, Ireland.

Vilar, D., Xu, J., D'Haro, L., Ney, H. 2006. Error analysis of statistical machine translation output. 5th International Conference on Language Resources and Evaluation, LREC'06, pp. 697–702.

Wagner, E. 1985. "Post-editing Systran – A Challenge for Commission Translators", Terminologie & Traduction, 1985-3.

Wagner, E. 1985. "Post-editing Systran – A Challenge for Commission Translators", Terminologie&Traduction, 1985-3.

Webb, N. M. 1988. Peer interaction and learning in small groups. Annual meeting of the American Educational Research Association, San Francisco.

Weisler, S.E., Milekic, S.P., Milekic, S. 2000. Theory of Language. MIT Press.

Whaley, C.P. 1978. Word-nonword classification time. Journal of Verbal Learning and Verbal Behaviour, 17, 143-154.

Williams, S. and Reiter, E. 2008. Generating basic skills reports for low-skilled readers. Journal of Natural Language Engineering.

Winerman, L. 2009. Crisis Communication. Nature, vol. 457, p 376

Wyner, A., Angelov, K., Barzdins, G., Damljanovic, D., Davis, B., Fuchs, N., Hoefler, S., Jones, K., Kaljurand, K., Kuhn, T., Luts, M., Pool, J., Rosner, M., Schwitter, R., and John Sowa. 2009. “On Controlled Natural Languages: Properties and Prospects.”. In Proceedings of CNL2009.

Zajic, D., Dorr, B., Lin, J., Monz, C. and Schwartz, R. 2005. A sentence-trimming approach to multi-document summarization. In Proceedings of DUC2005.

Zhang, W., and Yu, S. W. 1998. Construction of Controlled Chinese Lexicon. Proceedings of the Second International Workshop on Controlled Language Applications CLAW98. Pittsburgh, Pennsylvania: Language Technologies Institute, Carnegie Mellon University, pp. 159-173.

Zhu, Zh., Bernhard, D. and Gurevych, I. 2010. A monolingual tree-based translation model for sentence simplification. In Proceedings of the 23rd International Conference on Computational Linguistics COLING '10. Association for Computational Linguistics, Stroudsburg, PA, USA, 1353-1361.

Zipf G.K. 1949. Human Behavior and the Principle of Least Effort. Cambridge, Massachussetts: Addison-Wesley. p.1.

Андрейчин, Л. et al. 1993. Граматика на съвременния български език, том 2. Морфология, БАН, София.

Бояджиев, Т., Куцаров, И., Пенчев, Й. 1999. Съвременен български език. София: "Петър Берон".

Левенштейн, В.И. 1965. Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академий Наук СССР 163.4:845–848. Appeared in English as: V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10 1966.:707–710.

Appendix A: Previously Published Work

Some of the work described in this thesis has been published previously in the proceedings of peer-reviewed international conferences and workshops. All such research has been revised and significantly expanded before its inclusion in this thesis. This Appendix provides a short description of these articles (listed in chronological order) and explains their contribution to this thesis.

- **Temnikova, I.** and Orasan, C. (2009). *Post-editing Experiments with MT for a Controlled Language.* Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL), Besançon, France, July 1-3, 2009.

This article presents the evaluation of the impact of CLCM on manual and machine translation by calculating the time employed to manually translate the texts, the time employed to manually post-edit the texts, and the edit distance between the MT output texts and their post-edited versions. The evaluation presented in Sections 6.4.1, 6.4.2.1 and 6.4.2.2 is based on this article. Before inclusion in the thesis, the evaluation approach was improved by normalizing the time and edit distance per sentence length in characters, and the findings were revised on the basis of calculated statistical significance, which was lacking in the paper.

- **Temnikova, I.** and Margova, R. (2009). *Towards a Controlled Language in Crisis Management: The Case of Bulgarian.* Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL), Besançon, France, July 1-3, 2009.

The research described in this article was conducted in collaboration with one of the end-users of the MESSAGE project. It consisted of an experiment aiming to transfer the CLCM rules to the Bulgarian language. The contribution of the author of this thesis to this article consisted of adapting the MS kit to Bulgarian. Before inclusion in the thesis, the information about this research was significantly revised and expanded. The research described in this article was used as the basis for Section 4.3.4.

- **Temnikova, I.** (2010). *A Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment.* International Conference "Language Resources and Evaluation" (LREC2010), Valletta, Malta. May 17-23, 2010.

This article introduces to the research community the innovative evaluation approach of assessing the cognitive effort applied by post-editors to the MT output texts. It is described in Section 6.4.2.3. Before inclusion in the thesis, this research was significantly expanded in terms of hypotheses, motivations, related work, analysis of results, and discussion.

- **Temnikova, I.** (2011). *Establishing Implementation Priorities in Aiding Writers of Controlled Crisis Management Texts.* International Conference "Recent Advances in Natural Language Processing" (RANLP 2011), Hissar, Bulgaria. September 12-14, 2011.

This article presents an experiment aiming to investigate the concrete difficulties encountered by linguists while manually simplifying texts, and directions for future implementation priorities. The research presented in this article was used as a basis for Chapter 7, but has been significantly expanded in terms of additional analyses, results, findings, and conclusions.

- **Temnikova**, I., Orasan, C. and Mitkov, R. (2012). *CLCM - A Linguistic Resource for Effective Simplification of Instructions in the Crisis Management Domain and its Evaluations.* Accepted at the forthcoming International Conference "Language Resources and Evaluation" (LREC 2012), Istanbul, Turkey. May 21-27, 2012.

This article describes the CLCM linguistic resource in its final form, along with multi-perspective evaluations which support its usefulness. The main part of this article (the description of the controlled language resource) has been significantly expanded and enriched with multiple analyses in Chapter 4, while the evaluation sections contain only limited information about the most significant discoveries in Chapters 5, 6 and 7.

- **Temnikova, I.** and Cohen, K. B. (2012). *The Crisis Management Corpus and its Application to the Estimation of Crisis Management Communication Efficiency.* Accepted at the forthcoming workshop "Language Resources for Public Security Applications" at the International Conference "Language Resources and Evaluation" (LREC 2012), Istanbul, Turkey. May 27, 2012.

This article presents a new linguistic resource, namely the Crisis Management Corpus, which fills a gap in the availability of language resources in the domain of NLP for Crisis Management. The article contains very limited information about the corpus composition and its text complexity analysis; this has been significantly expanded in Chapter 3.

Appendix B: The Controlled Language for Crisis Management (CLCM) Guidelines[45]

45 Guidelines, developed during MESSAGE project, full title: Alert Messages and Protocols, project financed by the European Union (JLS/2007/CIPS/022). With the support of the Prevention, Preparedness and Consequence Management of Terrorism and other Security-related Risks Programme European Commission - Directorate-General Justice, Freedom and Security.

Table of Contents

Guidelines of the Controlled Language for Crisis Management (CLCM) for English.....424

Guidelines of the Controlled Language for Crisis Management (CLCM) for English

General Settings

- These guidelines address the writing and re-writing of easy to understand emergency instructions for the general (non-specialist) population. The guidelines can be easily adapted to other domains or document-types.
- The rules are divided according to:
  - the type of document they refer to
  - the part (title, section, conditions, instructions) of document they refer to
  - the type of rule

  Note: The parts vary according to the type of document
- There are the following types of rules:
  - **General** – describing the elements of a document, their order, and other presentation issues
  - **Formatting** – describing the formatting that should be used (indentation, fonts, etc.)
  - **Syntactic** – describing the syntactic restrictions
  - **Lexical -** describing the lexical restrictions
  - **Punctuation -** describing the punctuation restrictions
- Every rule is described by the following notation: "Dt_Pt_Rt_N"

Dt (Document type):	**Pt (Part type):**	**Rt (Rule type):**	**N (Number):**
Pr (protocol) **In** (instructions) **Ame** (alerts and messages) **PrNorm** (protocol for	**T** (title) **St** (title of a sub-section) **Cd** (condition) **I** (instruction) **L** (list)	**G** (General) **F** (Formatting) **S** (Syntactic) **P** (Punctuation) **L** (Lexical)	Rule No.

prevention/after an emergency) **PrDur** (protocol for during an emergency) **InNorm** (instructions for prevention/after an emergency) **InDur** (instructions for during an emergency) **InNormS** (instructions for prevention or after an emergency for specialists) **InDurS** (instructions for during an emergency for specialists) **InNormG** (instructions for prevention or after an emergency for the general audience) **InDurG** (instructions for during an emergency for the general audience) **AmeS** (alerts and messages for specialists) **AmeG** (alerts and messages for the general audience)	**Cm** (comment) **Mb** (message body)		

Example:

PrDurS_T_S_3 - the 3rd syntactic rule, regarding the titles of the protocols to follow during an emergency

Grammatical terms mini-dictionary /prototype/

Term	Definition	Examples
Part of speech	Part of speech refers to the terms by which we categorise words.	noun, verb, adjective.
Noun	A noun describes a 'thing'.	Ambulance, doctor
Verb	A verb describes an 'action'.	Jump, run, close
Active verb	Best explained by example. 'Someone did something' is active. 'Something was done' is passive.	The boy jumped.

Document type “Instructions”

General rules

In_G_00: Preserve the meaning and the information content of the original document.
In_G_01: This type of document is mainly aimed at the general audience (not specialists).
In_G_02: In some cases a specific audience (e.g. parents, children or University visitors) may be defined.
In_G_03: The document should contain easily identifiable parts: (title of the document, sections, titles of sections and sub-sections, compulsory actions to be always done, conditions, instructions, lists, comments)
In_G_04: The compulsory elements are in **bold**.
In_G_05: The optional elements are in **grey bold**.

In_G_06: The order of the elements is given in the list below:

- **a title**;
- **a note specifying a reference document(s)**
- **a note specifying the audience addressed**
- **a title of a sub-section of general situations**
- **the actions to be taken in general situation**
- **warnings**
- **title of a sub-sections of specific situations**
- **sub-sections about actions to be taken in specific situations**
- **every sub-section should contain:**

- any conditions under which the operations have to be performed
- **instructions**
- list of items

- comments, providing the reason, the aim or any other secondary but useful information about the operation to be performed

Guidelines for writing instructions for the general audience

In_G_07: Write the title according to the guidelines in Section "Guidelines for writing a title".

In_F_T_02: Jump 2 new lines after the title.

Title Instruction1 Instruction2 Instruction3	**Title** Instruction1 Instruction2 Instruction3

In_G_08: If there is a specific audience:
Write "Target audience: target audience".

Information for the parents	*Target audience: parents.*

Explanation: If there is no specific audience:
this element is optional.

In_G_09: If there are distinguished situations:

Identify the specific situations.
Divide the blocks of instructions regarding the specific situations in subsections.
Write first the most specific situation.
Write the next more general situation.
End with the most general situation.
Write a title for each subsection, following guidelines in Section “Guidelines for writing a title of a sub-section”.

	Remove the syringe Alteplase®.
	
	Connect an empty 10ml luer-lock syringe.
	If the patient is a newborn:
	  Draw 1ml of blood.
	If not:
	  Draw 5ml of blood.
	
	If the physician requests a hemodialysis:
	  Keep the drawn blood sample.
	If not:

	Throw the blood sample away.

In_G_10: If there are sub-situations:

Write the first sub-section title.
Jump 2 new lines before the first sub-section title. (In_St_G_01)
Write the subsection (conditions, instructions, comments)
Write the next sub-section titles.
Jump 2 new lines after each title of a sub-section. (In_St_G_02)
Jump 2 new lines after each subsection.

	Chemical attack **Attack outdoors** Go to the closest building. Take shelter quickly. Close all windows and doors. **Attack indoors** Follow chemical attack plans: Open windows. Breathe fresh air. Evacuate the building.

In_P_01: If you make reference to a specific document:
Put the document title in quotation marks.

	Consult technical file №3 “Aircraft engine maintenance”.

In_F_03: Separate each block of instructions with a new line.
In_F_04: Separate each group of conditions with a new line.

	Remove the syringe Alteplase®.
	
	Connect an empty 10ml luer-lock syringe.
	If the patient is a newborn:
	  Draw 1ml of blood.
	If not:
	  Draw 5ml of blood.
	
	If the physician requests a hemodialysis:
	  Keep the drawn blood sample.
	If not:
	  Throw the blood sample away.

In_G_11: Write in correct English.
In_G_12: Begin sentences with a capital letter.
In_G_13: Write only one piece of information per line.

Control severe bleeding by applying firm pressure to the wound using a clean, dry dressing and raise it above the level of the heart.	**How to control severe bleeding** Use a clean, dry dressing. Apply firm pressure to the wound. Raise the arm above the level of the heart.

In_G_14: Write the cardinal numbers in figures.

Cool with water for at least ten minutes.	Cool with water for at least 10 minutes.

In_G_15: Write the ordinal numerals fully in letters.

Hold the 1st pipe.	Hold the first pipe.

In_P_02: Put the proper punctuation sign at the end of each line, as defined for every document part (instructions, conditions, etc.).

In_P_03: Write a colon after the following elements:

- "If possible",
- "If not",
- "Perform the following actions simultaneously",
- "These are the instructions to follow",
- comments markers,

•conditions,
•instructions, followed by a list,
•elements of lists, followed by instructions.

	If you have any of the following symptoms: • difficulty breathing, • shortness of breath, • wheezing, • hoarseness, • high-pitched voice, • difficulty speaking, • chest pain, • chest tightness, • skin changes, • skin discharge, • increased pain where skin is burned , • stomach pain, • vomiting, • diarrhoea, • increased pain of exposed eyes, • discharge from exposed eyes :

	Call your doctor. OR Call the Emergency Department.

Example: Elements of a list followed by instructions.

In_S_01: Use only the allowed syntactical structures.
In_S_02: Avoid the forbidden syntactical structures on p.29.
In_S_03: Avoid demonstrative pronouns.

Take this bag.	Take the bag.

In_S_04: Avoid possessive pronouns.

Take his arm.	Take the arm of the patient.

In_S_05: Avoid personal pronouns.
Exception: The personal pronoun “You”.

If a person is unconscious: Give them mouth-to-mouth resuscitation.	If a person is unconscious: Give the person mouth-to-mouth resuscitation.

In_L_01: Choose the words in accordance with the lexical rules on p. 31.
In_L_02: Use only the words defined in the dictionary on p.33.
In_L_03: Avoid the forbidden lexical expressions on p.31.

In_L_04: If possible: Use the alternative expressions in the dictionary on p.33.

The patient suffered amnesia.	The patient suffered a memory loss.

In_L_05: Keep preposition and verb together in phrasal verbs.

Switch the lights *off*.	*Switch off* the lights.

Explanation: Preposition and verb separated by many words create difficulties for both non-native speakers and machine translation engines.

In_L_06: If possible:

Avoid acronyms and abbreviations.

If not:

Use only the acronyms and abbreviations pre-defined in the dictionary.

Contact the NPFS.	Contact the NPFS (**National Pandemic Flu Service**).

Explanation: Abbreviations can be ambiguous or unknown to non-native speakers.

In_S_06: Avoid passive voice.

Make sure 999 is called.	Call 999.

In_S_07: If possible:

Avoid negation.

Do not apply dry dressings.	Avoid contact with dry dressings.

Explanation: Negation is considered to be harder to understand than positive statements.

In_S_08: If possible:

Write 1 verb per sentence.

Wrap the affected part in cling film, do not apply dry dressings, keep the patient warm and call an ambulance.	Wrap the affected part in cling film. Avoid dry dressings. Keep the patient warm. Call an ambulance.

In_S_09: If possible:

Use Present Participle as an Adjective only.

Then bend his elbow while **keeping** the palm of his hand turned up.	Perform the **following** actions: Turn the palm of the victim up. Keep the palm of the victim turned up. Bend victim's elbow.

In_S_10: If you coordinate 2 elements:
Use the following conjunctions:
- OR,
- AND,
- NOR.

	Evacuate children AND elderly people.

In_S_11: If the elements you coordinate are verbs: Follow the rules for coordination of actions.

In_S_12: If you coordinate more than 2 elements:
Use a list.

Take with you home and car keys, a battery radio.	Take with you: • home keys, • car keys, • a battery radio.

In_S_13: Avoid omissions.

At the end, bandage.	At the end, bandage the wound.

In_S_14: If a preposition introduces 2 nouns:
Repeat the preposition.

If you are next to the exit or to a window:	If you are next to the exit or next to a window:

In_S_15: If an adjective determines more than one noun:
Repeat the adjective.

Spare clothes and blankets.	Spare cloths and spare blankets.

In_S_16: If an adjective and a complement determine the same noun:
Attach the adjective to the noun which it refers to.

The green dangerous products bin.	The green bin for dangerous products.

In_S_17: If 2 complements determine the same noun:
Repeat the noun.

Gas and electricity installations.	The gas installations and the electricity installations.

Guidelines for writing a title

In_T_G_01: Write a title that describes unequivocally only this document.
In_T_F_01: Use the following formatting:

3. Font style: **bold**.
4. Font size: at least 2 units bigger than the text
5. Alignment: >centred<

In_T_F_02: Jump 2 new lines after the title.
In_T_S_01: Use one of the following formulations:

- "How to + Imperative clause",
- "What to do in case of + NP without determiner",
- "What to do if + Conditional clause",
- "NP without determiner".

In_T_P_01: Avoid any punctuation signs at the end of the titles.

	How to avoid a fire **What to do in case of fire** **What to do if the patient is allergic** **Fire evacuation procedure**

In_T_S_02: If you use the formulation “How to + Imperative clause”:
Use positive form only.

How to not contaminate the patients	**How to protect the patients from contamination** **How to avoid patients' contamination**

Guidelines for writing a title of a section or a sub-section

*/Rules for writing section titles differ from **titles** only regarding the formatting/*

In_St_G_01: Jump two new lines before each title of a section.
In_St_F_01: Use the following formatting:

- Font style: **bold**.
- Font size: at least one unit bigger than text and one unit smaller than title.
- Alignment: left< .

In_St_G_02: Write a title that describes unequivocally only this section or sub-section.
In_St_F_02: Jump 2 new lines after the title of the section or the sub-section.
In_St_S_01: Use one of the following formulations:

- “How to + Imperative clause”,
- “What to do in case of + NP without determiner”,
- “What to do if + Conditional clause”,
- “NP without determiner”.

In_St_P_01: Avoid any punctuation signs at the end of the titles.
In_St_S_02: If you use the formulation “How to + Imperative clause”:
Use positive form only.

Guidelines for writing a condition

In_Cd_S_01: Use one of the following phrase structures:

- "If + Conditional clause + :",
- "In case of + NP without determiner + :",
- "As soon as + Conditional clause + :",
- "When + Conditional clause + :",

In_Cd_P_01: Put a colon at the end of conditions.

	If the patient suffers from Schizophrenia: In case of earthquake: As soon as you are safe: When help arrives: When you hear the fire alarm:

In_Cd_G_01: If you have 2 or more alternative conditions:
Start with the most specific one.
End with the most general one.

Connect an empty 10ml luer-lock syringe and draw 5ml of blood (1ml if it's a newborn)	Connect an empty 10ml luer-lock syringe. If the patient is a newborn: Draw 1ml of blood.

	If not: Draw 5ml of blood.

In_Cd_G_02: If you have 2 or more conditions
AND
If all conditions must be satisfied:
Write the first condition on the first line.
Write the conjunction “AND” on the second line.
Write the second condition on the third line.
Repeat the conjunction “AND” after each condition.

If a hurricane is approaching and you are at home:	If a hurricane is approaching AND If you are at home:

In_Cd_G_03: If you have to choose between 2 or more alternative conditions:
Write the first condition on the first line.
Write the conjunction “OR” on the second line.
Write the second condition on the third line.
Repeat the conjunction “OR” after each condition.

If the patient is a newborn or there is no previous medical history:	If the patient is a newborn OR

	If there is no previous medical history:

In_Cd_G_04: If you write 2 conditions,
AND
If the second condition excludes the first condition:
Write the first condition on the first line.
Write the block of instructions.
Write “If not:” on the next line.
Write the second condition.
Write the block of instructions.

Draw 5ml of blood (1ml if it's a newborn)	If the patient is a newborn: Draw 1ml of blood. If not: Draw 5ml of blood.

In_Cd_P_02: If there are more than 2 conditions that must be satisfied both,
OR
If you have to choose between 2 and more conditions:
Put a comma at the end of each condition.
Exception: the last condition.
Put a colon at the end of the last condition.

	In case of emergency, OR If the doctor is not there:

	Call for rescue.

Guidelines for writing an instruction

In_I_F_01: If the instruction is preceded by a condition:
Indentation: +1.
If not:
Indentation: 0.

	Connect an empty 10ml luer-lock syringe. (Instruction, indentation 0) If the patient is a newborn: (Condition) Draw 1ml of blood. (Instruction, indentation +1)

In_I_G_01: Write the instructions in a logical and chronological sequence.

Leave the building as quickly as possible. If possible: Turn off electricity, if you have time	If possible: Turn off electricity. If not: Leave the building as quickly as possible.

In_I_L_01: If possible:
Use discourse connectives (e.g. first, second, next, then, finally).

If you suspect there is an embedded object: Avoid pressing on the embedded object. Do the following actions simultaneously: - Press firmly on either side of the embedded object. - Build up padding around the embedded object. *Explanation: This needs to be done in order to avoid putting pressure on the object itself.* Finally, bandage the wound.	If you suspect there is an embedded object: Avoid pressing on the embedded object. Do the following actions simultaneously: - Press firmly on either side of the embedded object. - Build up padding around the embedded object. *Explanation: This needs to be done in order to avoid putting pressure on the object itself.* Bandage the wound.

Explanation: In this way the order of the instructions is clearer.

In_I_G_02: Use consecutive numbers for marking consecutive instructions.

If you suspect there is an embedded object: Avoid pressing on the embedded object. Do the following actions simultaneously: - Press firmly on either side of the embedded object. - Build up padding around the embedded object. *Explanation: This needs to be done in order to avoid putting pressure on the object itself.* Finally, bandage the wound.	If you suspect there is an embedded object: 1. Avoid pressing on the embedded object. 2. Do the following actions simultaneously: - Press firmly on either side of the embedded object. - Build up padding around the embedded object. *Explanation: This needs to be done in order to avoid putting pressure on the object itself.* 3. Finally, bandage the wound.

Explanation: In this way the order of the instructions is clearer.

In_I_G_03: Write only one action per instruction.

Wrap the affected part in cling film, do not apply dry dressings, keep the patient warm and call an ambulance.	Wrap the affected part in cling film. Avoid dry dressings. Keep the patient warm. Call an ambulance.

In_I_G_04: If you have 2 or more simultaneous actions:

Write the expression “Do the following actions simultaneously:”
Indent one tab to the right.
Write the instructions.
Avoid putting numbers.

Control severe bleeding by applying firm pressure to the wound using a clean, dry dressing and raise it above the level of the heart.	**How to control severe bleeding** Do the following actions simultaneously: Use a clean, dry dressing. Apply firm pressure to the wound. Raise the wound above the level of the heart.

In_I_G_05: If you have to choose between 2 or more alternative instructions:

Write the first instruction on the first line.
Write the conjunction “OR” on the second line.
Write the second instruction on the third line.
Repeat the conjunction “OR” after each line.
Do not write the conjunction “OR” after the last instruction.

Go outside or into a room with open windows. Leave the building as quickly as possible. Turn off electricity, if you have time.	Go outside. OR Go into a room with open windows.

In_I_G_06: If you have to choose between 2 instructions,
AND
If one of the instructions is preferable to the other one:
Write “If possible” on the first line.
Write the first instruction on the second line.
Write “If not” on the third line.
Write the second instruction on the forth line.

Leave the building as quickly as possible. Turn off electricity, if you have time.	If possible: Turn off electricity. If not: Leave the building as quickly as possible.

In_I_P_01: Put a dot at the end of an instruction.
Exception: When a list follows an instruction.
In_I_P_02: If a list follows an instruction:
Put a colon at the end of the instruction.

	Take with you: • a bottle of water,

	• a torch, • ready-to-eat food, • a mobile phone. Leave home immediately.

Guidelines for writing a list

In_Li_F_01: Use the following formatting:

- Font style: regular,
- Font size: same as instructions and conditions
- Alignment: left<,
- Bullets: dashes (bullets),
- Indentation: +1.

In_Li_G_01: If the list is not comprehensive:

Put “etc.” as last element of the list.

In_Li_P_01: Put a comma at the end of each element of the list.

Exception: the last element.

	Keep inflammable products far away from the following heat sources: • convector heaters, • light bulbs, • hot plates, • etc.

In_Li_P_02: Put a dot at the end of the last element of the list.
Exception: if some instructions follow the list.
In_Li_P_03: If instructions follow the list:
Put a colon at the end of the last element of the list.

	If the animal shows one of the following symptoms: • torpor, • progressive posterior paralysis, • anxiety, • aggressiveness: Take the animal to a veterinarian immediately.

In_Li_S_01: Use only NP with indefinite articles as elements of the list.

	Take with you: • a bottle of water, • a torch, • ready-to-eat food,

	• a mobile phone.

Guidelines for writing a comment

Comments mainly provide additional information. There are 3 types of comments (all optional):

- **Comment notes at the beginning of the document:**

  Target audience,
  Reference.

  They provide additional information about the document itself.

- **Comment notes following a condition or instructions:**
  Aim,
  Explanation,
  Exception,
  Definition,
  Example.

  They give additional secondary information about the reason why a condition or instruction has been formulated or additional explanations.

- **Warnings**
  They may be used to draw attention to a particular dangerous situation.

In_Cm_F_01: For Comments type 1 and 2 use the following formatting:

- Font style: *italic*,

- Font size: 1 unit smaller than instructions,
- Font colour: grey,
- Alignment: left<,
- Indentation: +1.

In_Cm_F_02: For Comments type 3 use the following formatting:

- Font style: regular,
- Font size: same as instructions,
- Font colour: red,
- Alignment: left<,
- Indentation: 0.

In_Cm_G_01: Use one of the following expressions:

- “Aim: ”,
- “Explanation: ”,
- “Exception: ”,
- “Ref.: ”,
- “Target audience: ”,
- “Warning: ”.

In_Cm_P_01: If the comment is a warning:
Put an exclamation mark at the end of the comment.
If not:
Put a dot at the end of each comment.

Warning: Specific situations exist.	Warning: Specific situations exist!

	Tap on pipes. *Aim: This will help rescuers to hear you.*

	Plan an escape route to follow at night. *Explanation: Most fire deaths and injuries occur while people are sleeping.* Avoid personal pronouns. *Exception: The personal pronoun "You".* *Target audience: parents.* *Ref.: www.mi5.gov.uk*

In_Cm_S_01: Use one of the following structures:

- Conditional clause,
- Imperative clause,
- Alphanumeric sequence,
- NP without determiner.
- NP with determiner.

Allowed syntactic structures

Conditional Clauses

Description:

In the grammar - dependent adverbial clauses with free positioning
(both initial and final placement are possible) regularly marked by
a subordinator indicating the relationship to the main clause.

Arg0 + Vconj + Arg1* + (Prep + Arg2)* + Mod* + Mod*

Neg = negation,
V = base form of the verb,
Vconj = conjugated form of the verb,
NP = noun phrase,
Arg0 = subject,
Arg1 = direct object,
Arg2 = indirect object,
Prep = preposition,
Mod = circumstance complements or adverbials
***** = optional element,
()* = a group of optional elements, but if one element of the group is present, the other ones are compulsory.

Examples:

C1	Arg0 + Vconj	If the patient dies:
C2	Arg0 + Vconj + Mod	If the patient dies during an operation:
C3	Arg0 + Vconj + Arg1	If the patient takes medicines:
C4	Arg0 + Vconj + Prep + Arg2	If the patient suffers from migraines:
C5	Arg0 + Vconj + Arg1 + Prep + Arg2	If the physician prescribes medicines to the patient:
C6	Arg0 + Vconj + Mod	If you live in London:
C7	Arg0 + Vconj + Arg1 + Mod	When you leave the children alone:
C8	Arg0 + Vconj + Arg1 + Mod + Mod	When you leave the children alone at home:

Imperative clauses

Description:

Characterized by the lack of subject, use of the base form of the verb,
absence of modal verbs and tense and aspect markers.
Urge to do something after the moment of speaking.

(Aux + Neg)* + V + Arg1* + (Prep + Arg2)* + Mod* + Mod* + Mod*

Neg = negation,
V = verb in base form,
Aux = auxiliary verb
Arg0 = subject,
Arg1 = direct object,
Arg2 = indirect object,
Prep = preposition,
Mod = circumstance complements or adverbials
* = an optional element
()* = an optional group of elements, but if one element of the group is present
the other ones are compulsory.

Examples:

I1	V	Go out.
I2	V+Mod	Go out immediately.

I3	Aux+Neg +V	Do not go out.
I4	V+Arg1	Close the doors.
I5	V + Arg1 + Prep + Arg2	Give an identity card to the children.
I6	V + Prep + Arg2	Go to the next stage.
I7	Aux + Neg + V + Arg1 + Mod	Do not leave the children alone.
I8	Aux + Neg + V + Arg1 + Mod + Mod	Do not leave the children alone at home.
I9	Aux + Neg + V + Arg1 + Mod + Mod + Mod	Do not leave the children alone at home without supervision.

Noun phrases with determiner

Det + Mod* + N + Mod*

N=Noun,
Det = Determiner
Mod = Modifier,
* = optional element.

Examples:

Dn1	Det + N	the patient
Dn2	Det + Mod + N	the internal staircase
Dn3	Det + N + Mod	a bottle of water
Dn4	Det + Mod + N + Mod	the government policy on terrorism

Noun phrases without determiner

/Usually used in the lists, titles, title of a sub-sections or in the conditional clause "In case of"/

Mod* + N + Mod*

N=Noun,
Mod = Modifier,
* = optional element.

Examples:

N1	N	burns
N2	Mod + N	identity card
N3	N + Mod	bottle of water
N4	Mod + N + Mod	government policy on terrorism

Forbidden syntactic structures /prototype/

Description:

Garden Path Sentences

Definition:

A **garden path sentence** is a grammatically correct sentence that starts in such a way that a reader's most likely interpretation is an incorrect one, luring them initially into an improper parse that then turns out to be a dead end. The "garden path" is a reference to the saying "to be led down the garden path", meaning "to be misled".

Examples:

The horse raced past the barn fell.
The old man the boat.
The man whistling tunes pianos.
The cotton clothing is made of grows in Mississippi.
The complex houses married and single soldiers and their families.
The author wrote the novel was likely to be a best-seller.
The man returned to his house was happy.
The government plans to raise taxes were defeated.

Non-restrictive Relative clauses

Non-Restrictive relative clauses are set off by commas.
Non-Restrictive relative clauses provide parenthetical, non-defining information.
Examples of non-restrictive relative clauses:

The liquid outer core**,** ***which*** might be compared to the outer two-thirds of an egg's yolk**,** reaches from 2,900 km to a depth of about 5,100 km.
The most famous of these early seismographs was invented by John Milne**,** ***who*** returned to Great Britain to establish the field of seismology.

Examples of restrictive relative clauses (ALLOWED but simplify):

Water pipes ***which*** came into the buildings through concrete walls were severed by the movement of the walls.
The gods ***who*** made the Earth gave it to a frog to carry on his back.

Unclear attachment

'Police seek orange attackers.'
= 'The police seek attackers who are orange'/ 'The police seek attackers who attacked an orange'/
'The police seek attackers who attacked with an orange'

'old men and women leave first'
- Bracketing ambiguity – not clear what the modifier refers to.
- ([old men] and women)/(old [men and women])

Ambiguous coordination

Description:

Avoid coordinating 2 actions with 'AND'.
Why? AND can mean both 'simultaneously' and 'one after another'. This can be sometimes crucial.
Example:
*'Center the steering wheel **and** lock in position.'*

Lexical rules /prototype/

Description:

General rules

1. Use only literal meaning.
2. Avoid idiomatic expressions.
3. Use concrete (instead of abstract) concepts.

Forbidden lexical expressions /prototype/

Description:

Ambiguous words

Description:

3. words with 2 or more meanings (e.g. 'right')

'After taking the right turn at the intersection, I..'

Pronouns

Description:

4. pronouns (e.g. 'it')

'Remove the bolt from the cover and slide it to the left.'

Vague quantifiers

Description:

- vague quantifiers (e.g. 'some', 'a few')

Technical terminology

Description:

Words with high age-of-acquisition

Description:

Figurative language

Description:

Figurative language is different from the literal language by assigning a non-literal meaning to ordinary words or expressions. The main types of Figurative language are: Metaphor, Metonymy, Idioms, Sarcasm and Humor.

A metaphor is a figure of speech which denotes one thing with the name of another. Example of a metaphor is: “That flat tire costs me an hour.” It should be replaced by the literal expression “That flat tire costs me very much.”

A metonymy is when one word is used in place of another word (the two words must be semantically related).
Examples or metonymies are:
“She is reading Shakespeare.” (Author's name for book)
“America doesn't want another Pearl Harbor” (Location's name for location)

Idioms are very often used metaphors. An example of an idiom is: “rains cats and dogs”.

Sarcasm is “a sharp, bitter, or cutting expression or remark”. Example of sarcasm is: ”Nice perfume. Must you marinate in it?”

Humour is a whole story which has a figurative meaning. An example of humour follows below:

“Why did the elephant sit on the marshmallow?” – “Because he didn't want to fall into the hot
“America doesn't want another Pearl Harbor” (Location's name for location)

Idioms are very often used metaphors. An example of an idiom is: “rains cats and dogs”.

Sarcasm is “a sharp, bitter, or cutting expression or remark”. Example of sarcasm is: ”Nice perfume. Must you marinate in it?”

Humour is a whole story which has a figurative meaning. An example of humour follows below:

“Why did the elephant sit on the marshmallow?” – “Because he didn't want to fall into the hot
chocolate.”

Orthographic neighbours

Description:

The number of existing words into which the current word can be transformed by changing one letter.

Example:

mine-N29: line, pine, mile...

Domain dictionary /prototype/

Description:

The domain dictionary is divided hierarchically in sub-areas, which can be further divided in sub-sub-areas.
The tables provide the term index, the dictionary from of the term, the term forms, the different existing term definitions, examples and optionally a preferred simple equivalent.

Sub-field: “First Aid Medical Terminology”

Description:

Index	Terms	Term Forms	Definitions	Examples	Alternative Sources: Simple English Wikipedia; Plain English Campaign

TFAMT_0001	shock	shocks (Pl), shocks (3 Sg.) shocked	*1. Noun*: a medical condition consisting of too little blood flow to the brain and other vital organs. Shock has many causes and in the later stages will usually result in a decreasing blood pressure. Shock is a severe condition that can be fatal. **Also Known As:** *Hypoperfusion*	haemorrhagic shock	
			2. *Noun*: an emotional state of mind, usually following a traumatic event such as a car accident or the loss of a loved one. This is probably the most common usage of the term.	After a terrible car accident, the driver is quiet and distracted. He is described as being in shock.	
			3. *Verb*: to deliver an electrical charge. In certain types of cardiac arrest, shocking the victim can allow the heart to restart and beat normally.	An automated external defibrillator (AED) shocks victims of cardiac arrest whose hearts are in ventricular fibrillation.	
TFAMT_0002	ambulance	Ambulances (Pl.)	A vehicle used to transport sick or injured people with medical needs. Ambulances can be cars, trucks, helicopters, boats, or airplanes. **Also Known As:** Mobile intensive care unit (MICU), rescue units, medical transport units.	Post emergency telephone numbers by phones (fire, police, **ambulance**, etc.).	
TFAMT_0003	amnesia		A condition in which memory is disturbed and/or lost		Loss of memory Source: Plain English

					Campaign
TFAMT_0004					
TFAMT_0005					
TFAMT_0006					

Step-by-step re-writing example

Original text:
If you suspect there is something embedded, take care not to press on the object. Instead press firmly on either side of the object and build up padding around it before bandaging to avoid putting pressure on the object itself.

/Passage, taken from www.redcross.org “How to treat severe bleeding”/

How to treat severe bleeding (write a short and a clear title)

If you suspect there is an embedded object:
(separate the condition and put it before the actions)
(avoid using unclear and ambiguous terms)

- Avoid pressing on the embedded object.
(the indentation helps understand that the actions should be executed only under this condition)
(the numbering of actions clarifies their execution order)
(re-write the negative sentence into a positive one, because it is more clear)
(avoid using unclear references, specify which object you are referring to)

- Do the following actions simultaneously:
("and" can mean both simultaneity and consecutiveness. It would be good to disambiguate this "and". If it means consecutiveness, just list the two actions one after another. If "and" means simultaneity, you have to specify this.)

    - Press firmly on either side of the embedded object.
    (avoid using pronouns as certain group of readers can't process them)

    - Build up padding around the embedded object.
    (specify which concrete object you mean in order to avoid ambiguity)

Explanation: In order to avoid putting pressure on the object itself.
(the indentation of the comment indicates that it is providing additional information about point 2)

- Bandage the wound.

(the indentation of the last action indicates its dependency of the main condition)
(introduce the omitted object in order to be more precise)

A re-written example

Before	**After**
If you suspect there is something embedded, take care not to press on the object. Instead press firmly on either side of the object and build up padding around it before bandaging to avoid putting pressure on the object itself. [Passage, taken from www.redcross.org "How To treat severe bleeding"]	**How to treat severe bleeding** If you suspect there is an embedded object: 1. Avoid pressing on the embedded object. 2. Do the following actions simultaneously: - Press firmly on either side of the embedded object. - Build up padding around the embedded object. *Explanation: This needs to be done in order to avoid putting pressure on the object itself.*

	3. Finally, bandage the wound.

Appendix C: Materials used for the online reading Comprehension experiment in Chapter 5

1. Complex-Simplified Pairs of texts

Complex Text 1 (ID 1 Set 0): Clean Your Home and Stop Mold - 160 words without title

Take out items that have soaked up water and that cannot be cleaned and dried. Fix water leaks. Use fans and dehumidifiers and open doors and windows to remove moisture. To remove mold, mix 1 cup of bleach in 1 gallon of water, wash the item with the bleach mixture, scrub rough surfaces with a stiff brush, rinse the item with clean water, then dry it or leave it to dry. Check and clean heating, ventilating, and air-conditioning systems before use. To clean hard surfaces that do not soak up water and that may have been in contact with floodwater, first wash with soap and clean water. Next disinfect with a mixture of 1 cup of bleach in 5 gallons of water. Then allow to air dry. Wear rubber boots, rubber gloves, and goggles when cleaning with bleach. Open windows and doors to get fresh air. Never mix bleach and ammonia. The fumes from the mixture could kill you.

==========

Simplified Text 1 (ID 2 Set 0): Clean Your Home and Stop Mold - 169 words without title (Subject 1)

How to clean your home and stop mold
Remove:
    items that have soaked up water
    AND
    items that cannot be cleaned and dried.
Fix water leaks.
To remove moisture:
    Use fans.
    Use dehumidifiers.
    AND
    Open doors AND windows.

Avoid mixing:
    Bleach AND ammonia.
Explanation: The fumes from the mixture could kill you.

If cleaning with bleach:
    Wear:
        rubber boots,
        rubber gloves,
        goggles.
    Open:
        windows AND doors.

Explanation: To get fresh air.

To remove mold:
Mix:
1 cup of bleach AND 1 gallon of water.
Wash the item with the bleach mixture.
Use a stiff brush to clean rough surfaces on the item.
Rinse the item with clean water.
Dry the item.
OR
Leave the item to dry.

Check
AND
Clean:
heating,
ventilating,
air-conditioning.

Ignore hard surfaces if:
they are soaked with water

To clean hard surfaces:
Mix:
soap AND clean water.
Use the mixture to wash items.
Mix:
1 cup of bleach AND 5 gallons of water.
Use the mixture to disinfect items.
Allow items to air dry.

==========

Complex Text 2 (ID 3 Set 1): After a Flood. Precautions When Returning to Your Home - 158 words without title

Electrical power and natural gas or propane tanks should be shut off to avoid fire, electrocution, or explosions. Try to return to your home during the daytime so that you do not have to use any lights. Use battery-powered flashlights and lanterns, rather than candles, gas lanterns, or torches. If you smell gas or suspect a leak, turn off the main gas valve, open all windows, and leave the house immediately. Notify the gas company or the police or fire departments or State Fire Marshal's office, and do not turn on the lights or do anything that could cause a spark. Do not return to the house until you are told it is safe to do so. Your electrical system may also be damaged. If you see frayed wiring or sparks, or if there is an odor of something burning but no visible fire, you should immediately shut off the electrical system at the circuit breaker.

==========

Simplified Text 2 (ID 4 Set 1): What to do when returning home after a flood - 140 words without title

Shut off:
Electrical power,
Natural gas tanks OR propane tanks.
Explanation: To avoid fire, electrocution or explosions.
If possible:
Return to your home during the daytime.
Explanation: To avoid using lights.
Avoid using:
candles,

gas lanterns,
torches.
Use battery-powered flashlights OR batter-powered lanterns.
If you smell gas,
OR
If you suspect a leak:
Turn off the main gas valve.
Open all windows.
Keep lights off.
Avoid doing anything that could cause a spark.
Leave the house immediately.
Notify one of the following:
gas company,
police,
fire department,
State Fire Marshal's office.
When you are told it is safe:
Return to the house.

If there is:
frayed wiring,
sparks,
OR
an odor of something burning AND no visible fire:
Immediately shut off the electrical system at the circuit breaker.
Explanation: There may be damage to the electrical system.
==========

Complex Text 3 (ID 5 Set 2): Facts About Personal Cleaning and Disposal of Contaminated Clothing - 159 words without title

As quickly as possible, wash any chemicals from your skin with large amounts of soap and water. Washing with soap and water will help protect you from any chemicals on your body. o If your eyes are burning or your vision is blurred, rinse your eyes with plain water for 10 to 15 minutes. If you wear contacts, remove them and put them with the contaminated clothing. Do not put the contacts back in your eyes (even if they are not disposable contacts). If you wear eyeglasses, wash them with soap and water. You can put your eyeglasses back on after you wash them. After you have washed yourself, place your clothing inside a plastic bag. Avoid touching contaminated areas of the clothing. If you can't avoid touching contaminated areas, or you aren't sure where the contaminated areas are, wear rubber gloves or put the clothing in the bag using tongs, tool handles, sticks, or similar objects.

==========

Simplified Text 3 (ID 6 Set 2): Subject 1 – 150 words

How to do personal cleaning

Mix soap and water.
Use large amounts of soap and water to wash chemicals from skin.
Explanation: Washing with soap and water will help protect you from any chemicals on your body.
If you wear contacts:
Remove the contacts.
Put the contacts with other contaminated items.
Avoid putting the contacts back in your eyes.

If your eyes are burning
OR
If your vision is blurred:
    Rinse your eyes with plain water for 10 to 15 minutes.
If you wear eyeglasses:
    Wash the eyeglasses with soap and water.
    Use the eyeglasses as normal.

How to dispose of contaminated clothes

Wash yourself.
Place your clothing inside a plastic bag.
Avoid touching contaminated areas of the clothing.
If you cannot avoid touching contaminated areas,
OR
If you are not sure where the contaminated areas are:
    Wear rubber gloves
    OR
    Put the clothing in the bag using:
        tongs,
        tool handles,
        sticks,
        etc.

==========

Complex Text 4 (ID 7 Set 3): Key Facts About Protecting Yourself After a Volcanic Eruption - 160 words without the title

You can do many things to protect yourself and your family after a volcanic eruption: Pay attention to warnings, and obey instructions from local authorities. For example, stay indoors until local health officials tell you it is safe to go outside. Listen to local news updates for information about air quality, drinking water, and roads. Turn off all heating and air conditioning units and fans, and close windows, doors, and fireplace and woodstove dampers to help keep ash and gases from getting into your house. Exposure to ash can harm your health, particularly the respiratory (breathing) tract. To protect yourself while you are outdoors or while you are cleaning up ash that has gotten indoors, use an N95 disposable respirator (“air purifying respirator”).If you don’t have an N-95 respirator, you can protect yourself by using a nuisance dust mask as a last resort, but you should stay outdoors for only short periods while dust is falling.

==========

Simplified Text 4 (ID 8 Set 3): How to protect yourself after a volcanic eruption - 169 words without title (Subject 1)

Pay attention to warnings.
Obey instructions from local authorities.
Example: Stay indoors until local health officials tell you it is safe to go outside.
Listen to local news updates for information about:
    air quality,
    drinking water,
    roads.
Turn off:
    heating,
    air conditioning units,
    fans.
Close:
    windows,

doors,
fireplace,
woodstove dampers.
Explanation: This helps keep ash and gases from getting into your house. Exposure to ash can harm your health, particularly the respiratory (breathing) tract.
If you are outdoors
OR
If you are cleaning up ash that has gotten indoors:
Use an N95 disposable respirator.
Follow directions for proper use of this respirator.
If you don't have an N-95 respirator:
If possible:
Avoid going outside.
If not:
Protect yourself by using a nuisance dust mask.
If dust is falling:
Stay outside for short periods only.

2. Introductory paragraphs to each text

Explanation: Nuisance dust masks can provide comfort and relief from exposure to relatively non-hazardous contaminants such as pollen, but they do not offer as much protection as an N-95 respirator.

Set 1: Text IDs 1 and 2:

Text 1:

Imagine that you are at home after a flood. Read the following instructions about how to clean your home safely. Read the instructions as fast as possible. You will be given a limited amount of time to read them. Try to remember as much as possible of the information, including the order of actions and key details. Remember to answer the questions according to what was written in the text rather than according to common sense!
Press 'Continue' when you are ready to start reading the instructions.

Set 2: Text IDs 3 and 4:

Thank you for completing the questions about Text 1! You can have a break before continuing, otherwise, get prepared for Text 2:

Imagine that you are outside after a flood. Read the following instructions about how to return safely home and what dangers to avoid. Read the instructions as fast as possible. You will be given a limited amount of time to read them. Try to remember as much as possible of the information, including the order of actions and key details. Remember to answer the questions according to what was written in the text rather than according to common sense!
Press 'Continue' when you are ready to start reading the instructions.

Set 3: Text IDs 5 and 6:

Thank you for completing the questions about Text 2! You can have a break before continuing, otherwise, get prepared for Text 3:

Imagine that you have come into contact with dangerous chemicals.
Read the following instructions about how to clean yourself and what to do with your clothes. Read the instructions as fast as possible. You will be given a limited amount of time to read them. Try to remember as much as possible of the information, including the order of actions and key details. Remember to answer the questions according to what was written in the text rather than according to common sense!
Press 'Continue' when you are ready to start reading the instructions.

Set 4: Text IDs 7 and 8:

Thank you for completing the questions about Text 3! You can have a break before continuing, otherwise, get prepared

for the last text (Text 4):

Imagine that you are in a situation of a volcanic eruption.
Read the following instructions about how to stay safe and what to beware of in such a situation. Read the instructions as fast as possible. You will be given a limited amount of time to read them. Try to remember as much as possible of the information, including the order of actions and key details. Remember to answer the questions according to what was written in the text rather than according to common sense!
Press 'Continue' when you are ready to start reading the instructions.

3. Welcome and Goodbye texts

Welcome text:

Welcome to the text comprehension experiment!

Thank you for agreeing to participate!

This is a very short and simple experiment (it should take you not longer than 15 minutes). You will be given 4 short texts, which contain instructions for emergency situations. You will have a minute and a half to read each text (you can press continue if you finish reading it earlier) . Try to remember as much as possible of the information, including the order of actions. After each text you will be asked to answer 5 multiple-choice questions about it. You will be allowed to take a break before starting the next text.

Begin experiment

Goodbye text:

If you have any questions, comments or advices about this experiment, please contact me at irina.temnikova@gmail.com.

4. E-mail and instructions to participants

Could you help me to evaluate my work in text simplification by participating in a simple on-line experiment that will take about 15 minutes of your time? You will be reading short texts about emergency situations and then answering questions about them. I would be very grateful if you could also forward it to your colleagues, friends or students.

All the instructions are provided at the link. What it is needed to be done is to read 4 short texts and answer the questions after them. The time to read the texts is limited as it imitates an emergency situation and the time for answering the questions is being measured, so please avoid getting distracted while doing it. After each text however the participant can make a break. The whole experiment doesn't take more than 15 minutes in total. The experiment requires entering some personal information which will be used only for statistics and will not be published as it is anywhere.

This is the link of the experiment, any feedback will be appreciated!

http://clg.wlv.ac.uk/demos/irina/

Thank you very much in advance,

5. Questions and Answers per Set

Notes:

- The correct answers for all questions are N. 1 (Marked with “0” while implementing the experiment).
- The order of the actions is given in the correct order.
- To each question an answer “I don't know” has been added.
- For the experiment the order of questions and the order of answers were randomized.

Set 1: Text 1 (Complex) and Text 2 (Simplified)	
Question number	**Text of the question and answers**
25	According to the text, which protective clothing do you need to wear when cleaning with bleach? • Rubber gloves, rubber boots and goggles. • Rubber hood, rubber gloves and rubber boots. • Rubber gloves, rubber hood and goggles. • Rubber gloves, rubber gas mask and rubber hood.
27	According to the text, what should you definitely avoid doing? 6. Mix ammonia and bleach. 7. Touch items that have soaked up water. 8. Mix bleach and clean water. 9. Touch items that have not soaked up water.
28	According to the text, when should you wear rubber boots and rubber gloves? • When cleaning with bleach. • When cleaning with water. • When cleaning hard surfaces. • When opening windows.
29	According to the text, why should you avoid mixing bleach and ammonia? • Breathing the fumes from the mix could be fatal. • The mixture of bleach and ammonia could explode. • The mix of ammonia and bleach could dissolve items. • The bleach would change the ammonia's features.
33	According to the text, to remove mold, in which order do you have to perform the following actions? Put a number into each box, (e.g. 1, 3, 4) or if you don't know the order, put "100" into "Don't know". • Wash the item with the bleach mix. • Rub rough surfaces with a stiff brush.

	• Rinse the item with clean water. • Dry it or leave it to dry.

Set 2: Text 3 (Complex) and Text 4 (Simplified)	
Question number	**Text of the question and answers**
30	According to the text, you should avoid using: • Torches, gas lanterns and candles. • Candles, gas and battery-powered lanterns. • Candles, lanterns and torches. • Candles, torches and battery-powered lanterns.
31	According to the text, if you suspect there has been a gas leak or you smell gas, in which order do you have to take these actions? Put a number into each box, (e.g. 1, 3, 4) or if you don't know the order, put "100" into "Don't know". • Turn off the main gas valve. • Open all windows. • Go out of the house. • Notify the authorities.
36	According to the text, you should return to the house: • If you are told it is safe to do so. • After calling the fire department. • To avoid fire, electrocution or explosions. • If you smell gas.
38	According to the text, you should immediately shut off the electrical system at the circuit breaker, because: • The electrical system may be broken. • Your house may explode. • Your house may be on fire. • You see natural gas or propane tanks.
51	According to the text, which kind of lights you should use? • Battery-powered flashlights or battery-powered lanterns. • Solar-powered flashlights or solar-powered lanterns. • Mains-powered flashlights or mains-powered lanterns. • Wind-up flashlights or wind-up lanterns.

Set 3: Text 5 (Complex) and Text 6 (Simplified)	
Question number	**Text of the question and answers**
40	According to the text, you should put the contaminated clothing in a plastic bag, using: • Rubber gloves, sticks, tongs, tool handles or similar. • Tongs, sticks, clothing, rubber gloves or similar. • Tool handles, tongs, sticks or leather gloves. • Hands, sticks, tool handles, tongs or similar.
41	According to the text, in order to clean yourself, in which order do you have to take these actions? Put a number into each box, (e.g. 1, 3, 4) or if you don't know the order, put "100" into "Don't know". • Wash yourself. • Put your clothing in a plastic bag. • Avoid touching contaminated clothing.
42	According to the text, wash chemicals from your skin with: • Lots of soap and water. • A little bit of soap and water. • Lots of plain water. • Brush and lots of soap.
43	According to the text, rinse your eyes with plain water for 10/15 minutes if: • Either your vision is blurred or your eyes are burning. • Either your eyes are burning or your eyes are itching. • Either your vision is blurred or your skin is itching. • Either your skin is itching or your eyes are burning.
44	According to the text, you should wash yourself with soap and water because: • It would remove any chemicals from your body. • It would protect your body from contamination. • It would clean your eyes and your contacts. • It would remove any dirt from your body.

Set 4: Text 7 (Complex) and Text 8 (Simplified)	
Question number	**Text of the question and answers**
45	According to the text, you should listen to the local news about: • Air quality, drinking water, roads. • Earthquakes, floods, roads. • Heating, air-conditioning units, fans. • Windows, doors, fireplaces.
48	According to the text, which is the best way to protect your breathing tract? • With an N-95 respirator.

	• With a P-95 oil proof mask. • With an R-95 mask. • With an N-99 respirator.
49	According to the text, use the appropriate respirator when: • You are outside or you are cleaning ash inside. • You are inside and you smell a gas leak. • You are outside and you are cleaning ash. • You are outside and you smell a gas leak.
50	According to the text, why do you have to close windows, doors, fireplace and woodstove dampers? • To help keep ash and gases from getting into your house. Exposure to ash can harm your health. • To protect you while you are outdoors or while you are cleaning up ash that has gotten indoors. • To help you to pay attention to warnings, and to obey instructions from local authorities. • To help you to listen to local news updates for information about air quality, drinking water, and roads.

Appendix D: Materials used for the Translation and Post-editing experiment in Chapter 6

1. Guidelines to participants

You have to translate and post-edit 2 texts. The texts are split into sentences. Some sentences have been automatically translated into Maltese- those you have to post-edit, and some are left in English- these you have to translate into Maltese. The text is presented in two columns, in the left are the original sentences in English, in the right are aligned the sentences you have to work on.

The experiment measures the time (in the sense it makes comparison between the time you spent on the first text and the time spent on the second), so please try to not get distracted and do it at once. (The whole thing should take you around 30 minutes). If you need to interrupt it and come back later you can "pause" the time and then "start" it again.

The texts won't be published so do not spend time on making them in an elaborated style. They have just to be written in a correct everyday language and be clear and understandable. Also do not spend too much time to look for the correct technical term, as they are ideally destinated to the general population (non specialists).

==========How to use the online application: (Please read until the end before starting, also it is useful if you look at the screenshots I'm attaching)

In order to start, you have to input your family name at the link:

http://clg.wlv.ac.uk/demos/postedit/index.php

Then please from the last drop-down menu choose Maltese and press "Display only texts with this target language". Then choose "Text1: Individual Preparedness. Nuclear Attack".

The application allows you to move from sentence to sentence, forwards and backwards by selecting from the drop-down menu ("next" or "previous") and pressing "save". (unfortunately if you have to move from the last back to the first sentence you have to go back one by one "previous"+"save"->"previous"+"save")

When you think you are done with this text, please select from the same drop-down menu "Finish" and then "Save".

A button, saying "Congratulations!!! Get statistics." will appear. Please press it and enter your levels of English and Maltese (I guess "advanced/native or advanced").

After it has given you your results, please press "Try another text" and perform the same described before this time for "Text2: How to find clean air very quickly" again from English into Maltese. Please pay attention to enter the same name you used for the previous one.

==========Now you can start the experiment :) Thanks very much in advance!

2. Alternating sentences for Spanish for the complex text

Original	Translation
Overarching Goal	Objetivo general Next Save Pause
Avoid radioactive fallout: evacuate the fallout zone quickly or, if not possible, seek best available shelter.	Avoid radioactive fallout: evacuate the fallout zone quickly or, if not possible, seek best available shelter.
Specific Actions	Acciones específicas
1. Move out of the path of the radioactive fallout cloud as quickly as possible (less than 10 minutes when in immediate blast zone) and then find medical care immediately.	1. Move out of the path of the radioactive fallout cloud as quickly as possible (less than 10 minutes when in immediate blast zone) and then find medical care immediately.
2. If it is not possible to move out of the path of the radioactive fallout cloud, take shelter as far underground as possible, or if underground shelter is not available, seek shelter in the upper floors of a multi story building.	2. Si no es posible salir de la trayectoria de la nube radiactiva, tome refugio subterráneo en la medida de lo posible, bajo tierra o si la vivienda no está disponible, buscar refugio en los pisos superiores de un edificio de varios.
3. Find ways to cover skin, nose, and mouth, if it does not impede either evacuating the fallout zone or taking shelter.	3. Find ways to cover skin, nose, and mouth, if it does not impede either evacuating the fallout zone or taking shelter.
4. Decontaminate as soon as possible, once protected from the fallout.	4. Descontaminar a la mayor brevedad posible, una vez protegido de la lluvia.
5. If outside the radioactive fallout area, still take shelter to avoid any residual radiation.	5. If outside the radioactive fallout area, still take shelter to avoid any residual radiation.

3. Guidelines to error annotators and evaluators

Task:

Analyse the MT translation, according to the following error classification. Indicate the Type of the error (e.g. 3.1.), plus any additional
information as specified below:

1. Missing Words in the generated sentence -
   - Indicate the missing word
   - Indicate if the missing word is essential for expressing the meaning of the sentence or it will just make the sentence grammatical,
   - Indicate the Part-of-Speech of the missing word
2. Bad word order – put the concerned words in this colour
   - Indicate whether in order to re-write it correctly you need to move only single words or whole phrases.
3. Incorrect words in the generated sentence -
   1. The incorrect word changes the meaning of the sentence

(indicate if the translated word is not in the right Part-of-Speech, e.g. “Overaching” is translated as a Verb (incorrect) instead of as an Adjective (correct))

   2. The incorrect word is actually a correct word in an incorrect form

(e.g. Plural instead of Singular, or a non-existing form e.g.“paroli” instead of 'parole' in Italian)

   3. Introduced extra word in the generated sentence
   4. Bad stylistic choice of words

(e.g. a repetition or a translation of a formal word with a familiar synonym of vice-versa)

5. Wrongly translated idiomatic expressions

4. Punctuation errors -

1. Missing punctuation

re-write the sentence with the inserted missing punctuation
indicate if it changes the meaning of the sentence

2. Wrong punctuation

indicate if it changes the meaning of the sentence

4. MT errors evaluation results for Spanish (segment)

Overarching Goal	Objetivo general	OK	
Avoid radioactive fallout: evacuate the fallout zone quickly or, if not possible, seek best available shelter.			
Specific Actions	**Acciones** específicas	3.1. **Acciones** –correct POS, correct form. Doesn't change the meaning	
1. Move out of the path of the radioactive fallout cloud as quickly as possible (less than 10 minutes when in immediate blast zone) and then find medical care immediately.			
2. If it is not possible to move out of the path of the radioactive fallout cloud, take shelter as far underground as possible, or if underground shelter is not available, seek shelter in the upper floors of a multi story building.	2. Si no es posible salir de la trayectoria de la nube radiactiva, **tome** refugio **subterráneo en la medida de lo posible, bajo tierra** o si **la vivienda** no está disponible, **buscar** refugio en los pisos superiores de un **edificio de varios**. → Si no es posible salir de la trayectoria de la nube radiactiva, **tome** refugio **en la medida de lo posible bajo tierra** o si **la vivienda subterráneo** no está	(1) 3.1. **tome** – incorrect word, correct POS, correct form. Doesn't change the meaning (2) 2.1 bad word order; necessary to move single words. Changes the meaning of the sentence (3) 4.2. wrong punctuation. Changes the meaning of the sentence (4) 2.2. Bad word order. Necessary to move whole phrases. Doesn't change the meaning of the sentence. (5) 3.1. **la vivienda** – wrong meaning, correct POS. changes the meaning (6) 3.2. **buscar** – doesn't change the meaning, wrong POS (infinitive, and it should be imperative)	

Appendix E: Materials used for the Text simplification experiment in Chapter 7

1. Instructions for the participants and assisting document

Subject: Text Simplification Task. Day (1 or 2?) Date:

Read the printed text, then simplify (re-write it) according to the listed below rules in the respective "-simplified" file. You are also allowed to consult the MESSAGE-CLCM Guide if you need to.
Measure the time needed for simplification. If you need to interrupt – stop measuring the time.

Text title:	Time:

1. Rules for discourse structure organisation at text level

-identify the separate situations
-group information regarding the specific situations in separate blocks
-jump two new lines after every specific situation block
-provide a unequivocal title for each specific situation block
-use the allowed formulations for the titles
-jump two new lines after each title

2. Rules for discourse structure organisation at paragraph level

-order instructions in logical and chronological order
-place conditions before instructions
-use standard word order
-use the suggested formulations for conditions
-if you coordinate two conditions - write one on one line, then "AND"or "OR" and the other one on the second line
-put the more specific conditions before the more general ones
-if there are two actions to be done simultaneously, write: "Do these two actions simultaneously:"
-order explanations, exceptions and other notes after instructions

3. Concrete linguistic realization rules

-put a colon after a condition
-write only one action per line
-replace technical terms with common synonyms
-replace idiomatic expressions with literal ones
-replace enumerations with vertical lists
-write the cardinal numbers in figures
-expand the abbreviations at their first occurrence
-avoid any pronouns (personal, possessive, demonstrative)
-avoid ambiguous words
-keep the preposition and the verb together in phrasal verbs

-replace passive with active voice
-try to avoid negative forms
-if a preposition/adjective refers to more than 1 noun, repeat the preposition/adjective next to each noun
-if more than 1 complement determine the same noun, repeat the noun
-put a comma after each element of a list, except of the last one (put a dot at the end of the last element of a list).

2. Simplifier questionnaire

Subject: Manual Text Simplification Questionnaire

Think about the way you simplified the texts and please reply to the following questions:

Could you think of what was most difficult for you while simplifying?

From the following list of instructions indicate which were the most difficult/easy to follow and which of them would simplify/speed up your job if done automatically:

1. Rules for discourse structure organisation at text level

-identify the separate situations
-group information regarding the specific situations in separate blocks
-jump two new lines after every specific situation block
-provide a unequivocal title for each specific situation block
-use the allowed formulations for the titles
-jump two new lines after each title

2. Rules for discourse structure organisation at paragraph level

-order instructions in logical and chronological order
-place conditions before instructions
-use standard word order
-use the suggested formulations for conditions
-if you coordinate two conditions - write one on one line, then “AND”or “OR” and the other one on the second line
-put the more specific conditions before the more general ones
-if there are two actions to be done simultaneously, write: “Do these two actions simultaneously:”
-order explanations, exceptions and other notes after instructions

3. Concrete linguistic realization rules

-put a colon after a condition
-write only one action per line
-replace technical terms with common synonyms
-replace idiomatic expressions with literal ones
-replace enumerations with vertical lists
-write the cardinal numbers in figures
-expand the abbreviations at their first occurrence
-avoid any pronouns (personal, possessive, demonstrative)

-avoid ambiguous words
-keep the preposition and the verb together in phrasal verbs
-replace passive with active voice
-try to avoid negative forms
-if a preposition/adjective refers to more than 1 noun, repeat the preposition/adjective next to each noun
-if more than 1 complement determine the same noun, repeat the noun
-put a comma after each element of a list, except of the last one (put a dot at the end of the last element of a list).

From the following proposed automatic simplifications indicate how much each would simplify and speed up your job.

1 – will not help at all
2 – to a certain extent
3 – very much

The text is presented to you with highlighted separate thematic situations.
The text is presented to you with highlighted acronyms and abbreviations.
The text is presented to you with highlighted pronouns.
The text is presented to you with highlighted passive voice.
The text is presented to you with highlighted negative phrases.
The text is presented to you with highlighted nouns to which a preposition refers.
The text is presented to you with highlighted nouns to which an adjective refers.
The text is presented to you with highlighted technical terms.
The text is presented to you with highlighted and underlined verbs.
The text is presented to you with highlighted beginning of conditions.
The text is presented to you with highlighted beginning of instructions.
The text is presented to you with highlighted beginning of explanations.
The text is presented to you with highlighted whole conditional expressions.
The text is presented to you with highlighted whole instructions.
The text is presented to you with highlighted whole explanations.
The text is presented to you with highlighted phrasal verbs in case the main verb and the preposition are split up.
The text is presented to you with ambiguous lexical terms highlighted.
The text is presented to you with ambiguous syntactic expressions highlighted.

3. Original texts used for simplification

Text 1:

Clean Your Home and Stop Mold

Take out items that have soaked up water and that cannot be cleaned and dried. Fix water leaks. Use fans and dehumidifiers and open doors and windows to remove moisture. To remove mold, mix 1 cup of bleach in 1 gallon of water, wash the item with the bleach mixture, scrub rough surfaces with a stiff brush, rinse the item with clean water, then dry it or leave it to dry. Check and clean heating, ventilating, and air-conditioning systems before use. To clean hard surfaces that do not soak up water and that may have been in contact with floodwater, first wash with soap and clean water. Next disinfect with a mixture of 1 cup of bleach in 5 gallons of water. Then allow to air dry. Wear rubber boots, rubber gloves, and goggles when cleaning with bleach. Open windows and doors to get fresh air. Never mix bleach and ammonia. The fumes from the mixture could kill you.

Text 2:

After a Flood
Precautions When Returning to Your Home

Electrical power and natural gas or propane tanks should be shut off to avoid fire, electrocution, or explosions. Try to return to your home during the daytime so that you do not have to use any lights. Use battery-powered flashlights and lanterns, rather than candles, gas lanterns, or torches. If you smell gas or suspect a leak, turn off the main gas valve, open all windows, and leave the house immediately. Notify the gas company or the police or fire departments or State Fire Marshal's office, and do not turn on the lights or do anything that could cause a spark. Do not return to the house until you are told it is safe to do so. Your electrical system may also be damaged. If you see frayed wiring or sparks, or if there is an odor of something burning but no visible fire, you should immediately shut off the electrical system at the circuit breaker. Avoid any downed power lines, particularly those in water. Avoid wading in standing water, which also may contain glass or metal fragments. You should consult your utility company about using electrical equipment, including power generators. Be aware that it is against the law and a violation of electrical codes to connect generators to your home's electrical circuits without the approved, automatic-interrupt devices. If a generator is on line when electrical service is restored, it can become a major fire hazard. In addition, the improper connection of a generator to your home's electrical circuits may endanger line workers helping to restore power in your area. All electrical equipment and appliances must be completely dry before returning them to service. It is advisable to have a certified electrician check these items if there is any question. Also, remember not to operate any gas-powered equipment indoors. (See also “Carbon Monoxide Poisoning” at www.bt.cdc.gov/disasters/carbonmonoxide.asp.) See also “Reentering Your Flooded Home” at www.bt.cdc.gov/disasters/mold/reenter.asp.

Cleanup
Walls, hard-surfaced floors, and many other household surfaces should be cleaned with soap and water and disinfected with a solution of 1 cup of bleach to five gallons of water. Be particularly careful to thoroughly disinfect surfaces that may come in contact with food, such as counter tops, pantry shelves, refrigerators, etc. Areas where small children play should also be carefully cleaned. Wash all linens and clothing in hot water, or dry clean them. For items that cannot be washed or dry cleaned, such as mattresses and upholstered furniture, air dry them in the sun and then spray them thoroughly with a disinfectant. Steam clean all carpeting. If there has been a backflow of sewage into the house, wear rubber boots and waterproof gloves during cleanup. Remove and discard contaminated household materials that cannot be disinfected, such as wallcoverings, cloth, rugs, and drywall. See also “Protect Yourself from Mold” at www.bt.cdc.gov/disasters/mold/protect.asp.

After a Flood

Immunizations
Outbreaks of communicable diseases after floods are unusual. However, the rates of diseases that were present before a flood may increase because of decreased sanitation or overcrowding among displaced persons. Increases in infectious diseases that were not present in the community before the flood are not usually a problem. If you receive a puncture wound or a wound contaminated with feces, soil, or saliva, have a doctor or health department determine whether a tetanus booster is necessary based on individual records. Specific recommendations for vaccinations should be made on a case-by-case basis, or as determined by local and state health departments.

Swiftly Flowing Water
If you enter swiftly flowing water, you risk drowning -- regardless of your ability to swim. Swiftly moving shallow water can be deadly, and even shallow standing water can be dangerous for small children. Cars or other vehicles do not provide adequate protection from flood waters. Cars can be swept away or may break down in moving water.

Chemical Hazards
Use extreme caution when returning to your area after a flood. Be aware of potential chemical hazards you may encounter during flood recovery. Flood waters may have buried or moved hazardous chemical containers of solvents or other industrial chemicals from their normal storage places. If any propane tanks (whether 20-lb. tanks from a gas grill or household propane tanks) are discovered, do not attempt to move them yourself. These represent a very real danger of fire or explosion, and if any are found, police or fire departments or your State Fire Marshal's office should be contacted immediately. Car batteries, even those in flood water, may still contain an electrical charge and should be removed with extreme caution by using insulated gloves. Avoid coming in contact with any acid that may have spilled from a damaged car battery.

For more information, visit www.bt.cdc.gov or call CDC at 800-CDC-INFO (English and Spanish) or 888-232-6348 (TTY).

Text 3:

Facts About Personal Cleaning and Disposal of Contaminated Clothing

As quickly as possible, wash any chemicals from your skin with large amounts of soap and water. Washing with soap and water will help protect you from any chemicals on your body. o If your eyes are burning or your vision is blurred, rinse your eyes with plain water for 10 to 15 minutes. If you wear contacts, remove them and put them with the contaminated clothing. Do not put the contacts back in your eyes (even if they are not disposable contacts). If you wear eyeglasses, wash them with soap and water. You can put your eyeglasses back on after you wash them. Disposing of your clothes: o After you have washed yourself, place your clothing inside a plastic bag. Avoid touching contaminated areas of the clothing. If you can't avoid touching contaminated areas, or you aren't sure where the contaminated areas are, wear rubber gloves or put the clothing in the bag using tongs, tool handles, sticks, or similar objects. Anything that touches the contaminated clothing should also be placed in the bag. If you wear contacts, put them in the plastic bag, too. o Seal the bag, and then seal that bag inside another plastic bag. Disposing of your clothing in this way will help protect you and other people from any chemicals that might be on your clothes. o When the local or state health department or emergency personnel arrive, tell them what you did with your clothes. The health department or emergency personnel will arrange for further disposal. Do not handle the plastic bags yourself.

Text 4:

FACT SHEET

Key Facts About Protecting Yourself After a Volcanic Eruption

You can do many things to protect yourself and your family after a volcanic eruption: • • • • Pay attention to warnings, and obey instructions from local authorities. For example, stay indoors until local health officials tell you it is safe to go outside. Listen to local news updates for information about air quality, drinking water, and roads. Turn off all heating and air conditioning units and fans, and close windows, doors, and fireplace and woodstove dampers to help keep ash and gases from getting into your house. Exposure to ash can harm your health, particularly the respiratory (breathing) tract. To protect yourself while you are outdoors or while you are cleaning up ash that has gotten indoors, use an N95 disposable respirator (also known as an “air purifying respirator”). N-95 respirators can be purchased at businesses such as hardware stores. It is important to follow directions for proper use of this respirator. For more information, see “NIOSH-Approved Disposable Particulate Respirators (Filtering Facepieces)” (www.cdc.gov/niosh/npptl/topics/respirators/disp_part). If you don’t have an N-95 respirator, you can protect yourself by using a nuisance dust mask as a last resort, but you should stay outdoors for only short periods while dust is falling. Nuisance dust masks can provide comfort and relief from exposure to relatively non-hazardous contaminants such as pollen, but they do not offer as much protection as an N-95 respirator. Stay away from ashfall areas, if possible. Avoid contact with ash as much as you can. Keep your skin covered to avoid irritation from contact with ash. Wear goggles to protect your eyes from ash. Do not travel unless you have to. Driving in ash is hazardous to your health and your car. Driving will stir up more ash that can clog engines and stall vehicles. Replace disposable furnace filters or clean permanent furnace filters frequently. If your drinking water has ash in it, use another source of drinking water, such as purchased bottled water, until your water can be tested. Clear roofs of ash. Ash is very heavy and can cause buildings to collapse. Be very cautious when working on a roof. Ash can be slippery and make it easy to fall. Information about injuries and mass trauma events can be found in “Injuries and Mass Trauma Events: Information for the Public” (www.bt.cdc.gov/masstrauma/injuriespub.asp).

• • • • • •

Volcanic eruptions may result in floods, landslides and mudslides, power outages, and wildfires. For information on protecting yourself against these hazards, visit the following: • • • Earthquakes: www.bt.cdc.gov/disasters/earthquakes Includes information on preparing for, surviving, and recovering from an earthquake. Floods: www.bt.cdc.gov/disasters/floods Includes information on making sure food and water are safe, cleaning up, and emergency supplies. Landslides and mudslides: www.bt.cdc.gov/disasters/landslides.asp Includes information on protective measures to take before, during, and after a landslide or debris flow. March 9, 2005
Page 1 of 2

Key Facts About Protecting Yourself After a Volcanic Eruption (continued from previous page) • • Power outages: www.bt.cdc.gov/poweroutage
Includes information on carbon monoxide poisoning, alternative heat and energy sources, downed power lines, and food and water safety.

Wildfires: www.bt.cdc.gov/firesafety Includes information on smoke inhalation and other wildfire hazards.

Sources
For more information on volcanoes and health, see the following sources: • • • American Red Cross o “Volcano”: www.redcross.org/services/disaster/0,1082,0_593_,00.html Federal Emergency Management Agency o “Fact Sheet: Volcanoes”: www.fema.gov/hazards/volcanoes/volcanof.shtm o “Volcanoes: Are You Ready?”: www.fema.gov/areyouready/volcanoes.shtm U.S. Geological Survey o “What To Do if a Volcano Erupts”: http://vulcan.wr.usgs.gov/Hazards/Safety/framework.html o “Volcanic Ash and Mudflows”: http://vulcan.wr.usgs.gov/Hazards/Safety/what_to_do_EIB.html o “Volcanic Gas”: http://vulcan.wr.usgs.gov/Projects/Emissions/vgas_fsheet.html Washington State Department of Health o “Volcanoes”: www.doh.wa.gov/phepr/handbook/volcano.htm (also available in Spanish: www.doh.wa.gov/phepr/handbook/spanish_pdf/volcan_spanish.pdf)

For more information, visit www.bt.cdc.gov/disasters/volcanoes, or call CDC at 800-CDC-INFO (English and Spanish) or 888-232-6348 (TTY). March 9, 2005